Institute of Biotechnology
Research Program in Structural and Quantitative Biology
University of Helsinki

Department of Physics
Faculty of Science
University of Helsinki
Finland

Structural and biochemical studies of a post-translational modification platform in early Eukaryota

Valerio Chiarini

To be presented for public discussion with the permission of the Faculty of Science of the University of Helsinki, on the 28th of October 2021, at 10 o'clock. The defence is open for audience through remote access.

Helsinki 2021

ISBN 978-951-51-7577-9 (paperback)
ISBN 978-951-51-7578-6 (PDF)

Supervisor

Dr. Vivek Sharma
Department of Physics
University of Helsinki
Finland

Reviewers

Professor Lari Lehtiö
Biocenter Oulu and Faculty of Biochemistry and Molecular Medicine
University of Oulu
Finland

Professor Antonio Macchiarulo
Department of Pharmaceutical Sciences
Università degli studi di Perugia
Italy

Opponent

Dr. Paola Storici
Senior Scientist in Structural Biology, Head of Protein Facility
ELETTRA Synchrotron of Trieste
Italy

Custos

Professor Adrian Goldman
Molecular and Integrative Biosciences Research Programme
Doctoral Programme Brain & Mind
Doctoral Programme in Integrative Life Sciences
Finland

ACKNOWLEDGEMENTS

The content of this thesis is the result of the work carried out between 2015 and 2020 at the University of Helsinki (2015-2017 Institute of Biotechnology and 2017-2020 Department of Physics) in collaboration with “La Sapienza” University of Rome.

In terms of the former institute, I want to express my gratitude to those who have helped me out during one of the most challenging times of my life, starting from the members of my committee, Tommi Kajander and Hannu Maaheimo, for their consideration and support of which I always benefited. For the same reasons, highly regarded is Helena Tossavainen, co-author, and good friend. Know that your office was always an antidote for lots of things, and not just for me.

To all the members of Hideo Iwai’s group, in particular Annika, Britta, Fernando, Tuuli, Sandra, Sesilija, Taru and Harri: I have treasured memories of you all. Thank you for your help, suggestions, chats, and friendship. A special thanks goes to Tuomas for reaching out to me and offering a bright side for me to hold on to when it was most needed. I will never forget your kindness, nor that of your wife and your daughters. I also want to thank the Integrative Life Science (ILS) doctoral program for granting me the opportunity of pursuing the PhD in Finland.

Regarding the department of Physics, I want to thank all the members of Ilpo Vattulainen group, which have welcomed me and walked me through the secrets of MD, in particular Amina Djurabekova and Fabio Lolicato.

Also, the Academy of Finland and CSC are acknowledged for the resources provided in the project.

My deepest gratitude goes to Vivek Sharma, my second and only supervisor. Since you took the burden of dedicating time to me and my scientific goals, not one time I felt alone, undermined, or a saboteur, but I grew strong in my resilience and eager to achieve the results which I initially set out to achieve. By seeing the person and the student as one, you proved that supervision has nothing to do with babysitting.

Beyond the working environment, I want to thank my friends Roland, Julia, Blanca, Loïc and Francesca for the warmth of your friendship and the loving moments that we shared together, might they have been the cheese-nights, the adventure in Kilpisjärvi or the endless nights spent playing Resistance. On a separate note: thanks to Fabien. You have been a wonderful friend and roommate and whatever of what we occasionally had to endure of each-other, I would do again considering everything else we did together. Umair: you will always have a special place in my heart, but you already know that. Heini and Pasi: you were the unexpected. I did not expect to grow such affection towards someone, but you guys.....actually, in the end it might have just been about the food the whole time.

At “La Sapienza” University, I wish to thank first and foremost, Gianni Colotti and Andrea Ilari for all the work I could conduct in their lab and because I would not be writing these lines if it weren’t for them as well. Both I consider mentors but before that, dear friends. In particular, I want to thank Gianni for believing in me since before my PhD and Andrea as well, along with his patience in teaching me the basis of crystallography,

for which I equally want to thank Annarita Fiorillo. I look forward to working on Maltodextrin together again in the future.

Another special thanks goes to the best lab-peers I have ever had: Ilaria, Theo and Dr. Mocci. My fortuitous arrival in Rome could not have been "scheduled" better in time for us to be altogether. You always backed me up, often by just saying "It's okke, but it's now or never" and you never sparred criticisms. That makes you fine scientists and great friends at the same time. A great role had my family and in particular my father who, on that 14 November, worriedly watched me taking the same plane he later on took several times, in order to visit me as well for enjoying the awe that he developed for Finland. Thank you for experiencing this adventure with me, mum would have loved it just as much. For being there, always, I thank my friends in Rome: (old ones) Fabrizio, Irene, Michele, Riccardo, Max, Giulia (and new-old ones) Martina. They silently endured all my PhD tales and dramas.....thank you for the patience, I love you!

Finally, I want to dedicate my biggest thanks to my better half, Giulia. The toll that a similar experience would have taken from most couples, this PhD could not take from us. Not only did you wait for me, but you have always been there, no matter the distance, the obstacles, or the challenges, you were there, wishing for my happiness and cheering at my accomplishments. Had I been able to choose, you never would have had to see me as you did, at my worst and lowest, and yet, despite that, you never left my side. Instead, you filled the emptiness that more than once I felt inside myself. You were, and still are, my pillar. In hindsight, while I am happy to have pursued a PhD as much as I'm happy to be with you, I'm glad that the first is over and the latter is not. I love you with all my heart.

"…drive away from Derry, from memory…but not from desire. That stays, the bright cameo of all we were and all we believed as children, all that shone in our eyes even when we were lost and the wind blew in the night.
Drive away and try to keep smiling. Get a little rock and roll on the radio and go toward all the life there is with all the courage you can and all the belief you can muster.
Be true, be brave, stand.

All the rest is darkness"

Stephen King 彡

INDEX

LIST OF ORIGINAL PUBLICATIONS

This thesis was written using the data contained in the following papers indicated in the main text by Roman numbering.

(I) Chiarini V, Tossavainen H, Sharma V, Colotti G (2019). NMR Structure of a Non-Conjugatable, ADP-ribosylation Associated, Ubiquitin-Like Domain from Tetrahymena Thermophila Polyubiquitin Locus. *BBA General Subject* **4:**749-759

(II) Chiarini V, Fiorillo A, Camerini S, Crescenzi M, Nakamura S, Battista T, Guidoni L, Colotti G, Ilari A. Structural basis of ubiquitination mediated by protein splicing in early Eukarya. *BBA General Subject* 1865 129844.

Author contribution:

V.C. conceived the project, purified the proteins, took care of NMR studies apart from spectral acquisition and performed the bioinformatics analysis and simulations with the counsel of V.S. V.C. wrote the manuscript along with G.C., H.T. and V.S.

II V.C. conceived the project, designed the experiments, and carried out protein expression and purification with the cooperation of T.B. Crystallization was performed by V.C and A.F while structure determination was carried out by V.C, supervised by A.I. S.C and M.C took care of the MS characterization. S.N. and L.G. performed Molecular Dynamics analysis. V.C, A.I, and G.C wrote the manuscript.

ABBREVIATIONS

Amp	Ampicilline
ART	ADP-ribosyl-transferase domain
ATP	Adenosine triphosphate
BIL	Bacterial intein like
BUBL	BIL-ubiquitin-like
COSY	Correlation spectroscopy
CPS	Conditional protein splicing
CSI	Chemical-shift index
D_2O	Deuterated water
DTT	Dithiotreithol
E.coli	*Escherichia coli*
EMP	Ethyl-mercuryl-phosphate
FNB	Frame not bound
GST	Glutathione-S-transferase
GTP	Guanosine-5'-triphosphate
HCD	Higher-energy collisional dissociation
HINT	Hedgehog-intein
HR	Homologous recombination
HSQC	Heteronuclear single quantum coherence
IMAC	Immobilized metal ion affinity chromatography
IPTG	Isopropyl-b-D-1-thiogalactopyranoside
IR	Infra-red
Kan	Kanamycine
LB	Luria Bertani
LCF	Loss contact per frame
MD	Molecular Dynamic
MR	Molecular replacement
NMR	Nuclear magnetic resonance
NOESY	Nuclear overhauser enhancement spectroscopy

OD	Optical density
P.horikoshii	*Pyrococcus horikoshii*
POI	Protein of interest
P.tetraurelia	*Paramecium tetraurelia*
PTM	Post translational modification
PTS	Protein trans splicing
R.filosa	*Reticulomyxa filosa*
RMSD	Root main square deviation
RMSF	Root main square fluctuation
RU	Resonance unit
SAD	Single-wavelenght anomalous dispersion
SPR	Surface plasmon resonance
TOCSY	Total correlation spectroscopy
T.thermophila	*Tetrahymena thermophila*
T.sibericus	*Thermococcus sibericus*
UBD	Ubiquitin binding domain
UBL	Ubiquitin like domain
UIM	Ubiquitin interacting motif
Ulp1	Ubiquitin-like protease 1

ABSTRACT.

"Structure is more conserved than sequence". This unanimously accepted concept, which states that two proteins with low sequence similarity (~30%) still feature the same fold, holds true only regarding the correlation between the sequence and the secondary structure elements (mainly helices and sheets) thanks to which an ordinate, three-dimensional arrangement is formed (i.e. in globular domains). For this reason, proteins lacking defined secondary motifs usually display no tertiary structure and are therefore designated as "disordered". As such, their sequence offers no information about their structure. Interestingly, multi-domain proteins face a similar problem. That is because, although the sequence is informative about the structure of the individual globular domains, their three-dimensional arrangement depends on the domains' surfaces and degree of freedom upon folding, which cannot be drawn from the sequence. That also means that, while (as stated before) the fold of the individual domains is mainly resistant to mutations, their relative position may be easily altered by them. That is the reason why multi-domain proteins often exert the biological functions of adaptors or scaffold elements instead of performing catalytic activity, for which the formation of an active site at the domain-domain interface is usually required.

Due to the limitations of deriving a biological function from sequences of multi-domain proteins, this study structurally and functionally characterizes a three-domain protein (BUBL), formed by an intein flanked by two ubiquitin-like-domains (ubl). It is here demonstrated that BUBL exerts both catalytic and decoying functions as it can conjugate by protein

splicing one of the three domains (N-ubl) either to itself or to a separate protein (*Tth*Ras GTPase) which is specifically lured by the C-ubl (baiting function). Resulting non-canonical ubiquitination, occurring in a single, concerted step and without energy consumption, is a representative example of how molecular evolution can produce the same biological goal by subverting the structural conservation normally required for its achievement. In the case presented here, two different post translational modifications (ubiquitination and protein splicing) are shown to functionally coexist in a unique combination forming a post translational platform, originated by the serendipitous insertion of the intein domain which, in most cases, is biologically inconsequential.

This thesis discusses the proposed scientific hypothesis by using bioinformatics, modelling and simulation approaches in combination with experimental techniques of biochemistry and structural biology.

1. INTRODUCTION

1.1 Post-Translational modifications: the right slang for the right venue.

If the relationship between the different levels of cellular regulation was to be explained by a figurative example, it could be represented with the series of choices made in order to attend an event. The first level of regulation concerns the connotation of a person. Mild adjustment such as shaving or having the hair done can be important but won't change the overall aspect of a person (like posture, facial expressions, etc), and most importantly, with respect to the event, are irreversible choices which must be done beforehand. Analogously, although the DNA content doesn't change between cells of the same organism, the way it is packed can differentiate a cell into many. Such as "shaving", differentiation is also (most of the time) irreversible. The second level, which is probably the most important one, is the transcriptional regulation of the genomic information. It provides for the expression of a subset of genes needed during different stages of the cell cycle or in response to exogenous signals. As transcription depends on what the cell is about to do, it can be compared to the dress-code which fits most the "venue" of the event. In this case, as clothing items are chosen to serve multiple purposes (showing elegance while providing warmth and comfort), transcription also generates different molecules (mRNA, rRNA, tRNA, long non-coding RNA, etc.), which together cooperate in shaping the new cellular conditions.

Compared to differentiation, transcription is a faster process whose constant regulation responds to tiny variations of incoming signals. The third and the last level, comprises the post-translational modification of proteins. As soon as proteins became object of study, it was immediately understood that their chemical composition was more diverse and complex than expected (Ambrogelly, Palioura, and Söll 2007). Differently from the DNA, whose zipper-like structure fit the function of bearer of the sensible genetic information, proteins were soon recognized as "enactors" of that information on a stage of limitless different contexts. The ability to perform according to the context was initially achieved by "translating" the four-base alphabet into a larger, more heterogeneous code of twenty amino acids. Because of the chemistry of these new bricks, different sequences of amino acids could generate different molecular structures, each one of them optimized to carry out a specific function. Sadly, the number of "enactors" produced just by translation was limited by definition to the number of sequences available in the genomic pool, including isoforms produced by alternative splicing. Meaning that, in order to provide a set of molecular machineries able to function under any circumstances, organisms had to develop an infinitively long DNA, containing specific sequences for each thinkable scenario. While this option is unfeasible, given the infinite amount of energy required to build such string, nature solved the problem by expanding the inventory of chemical groups which proteins can be composed of. This strategy achieves the goal of generating a large subset of almost identical proteins from a single coding sequence, where each protein differs from another one only for small chemical groups. Each protein can therefore be post-translationally modified in

several fashions, which reflect its functional state in space and time. The most powerful aspects of post-translational modifications (PTMs) is that most of them are reversible and are used in combination with others, hence elevating the number of protein-protein interactions to the power of the modifications each protein can undergo in time and space. For example, protein maturation, folding and compartmentalization is dictated by long chains of saccharides attached on either asparagine (N-glycosylation, endoplasmatic reticulum) or serine/threonine (O-glycosylation, Golgi apparatus) residues (Reily et al. 2019). Such chains are progressively trimmed as the protein proceeds in its maturation or re-elongated if misfolded proteins need to interact again with molecular chaperones. Analogously, proteins like kinases, which initiate molecular signalling cascades, could not integrate exogenous inputs if located away from the membrane. For this purpose, proteins are modified with lipid anchors which keep them attached to the membrane (Saha, Anilkumar, and Mayor 2016). It's easy to understand that such system allows to regulate the whole proteome of several fold more rapidly and specifically than transcription, which regulates protein production like an on/off switch. If the latter is compared to the dress-code of the event, PTMs can be thought as the conversational skills which allow to dynamically interact with the guests in real time. The more diverse the vocabulary, the slang and the conversation topics are, the more easily the interaction can be initiated or ended. Nowadays, more than 200 different PTMs have been identified over different classes of proteins (Fig 1). Interestingly, because PTMs exponentially increase the protein diversity, it was observed that the number of proteins responsible for PTMs correlates with the complexity of

the organisms. In humans, they represent up to 5% of the entire proteome (Aebersold et al. 2018).

Type	Function	PTMs
ENZYMATIC in vivo	Hydrophobic groups	Miristoylation, Palmitoylation, Isoprenylation (2 variants), Glipyation, Lipoylation
	Cofactors Enyzme cathalisis	FMN / FAD, Heme, Shiff Base, Phosphopantetheinylation
	Modification of translation factors	Diphtamide formation, Hypusine, Ethanolamine phosphoglycerol, Beta-Lysine
	Small chemical groups	Acetylation, Alkylation, C-ter amidation, Butyrilation, Gamma-carboxylation, Glycosylation, Malonilation, Hydroxylation, Iodination, ADP-ribosylation, Phosphorylation, Adenylylation, Uridylylation, Propionylation, Pyroglutamate formation, S-glutathionylation, S-nitrosylation, S-sulphenylation, S-sulphinylation S-sulphonylation, Succinylation, Sulfation
	Protein PTM	ISGylation, Ubiquitination, Sumoylation, Neddylation, Pupylation
	Aminoacid modification	Citrullination, Deamination, Eliminylation, Racemization
NON-ENZYMATIC in vivo		Cystine formation, Glycation, Isoaspartate formation, Carbamylation, Carbonylation, Isopeptide bond formation

Fig 1. Summary table of known post-translational modification (PTMs) divided as enzymatic and non-enzymatic. Enzymatic PTMs are further divided by subclasses indicating the chemical modification addressed to the target protein. Non-enzymatic isopeptide bond formation corresponds to Asp isomerization process, not to be confused with the isopeptide formation occurring during the ubiquitination process

1.2 Ubiquitination: a protein based PTM

While most PTMs are composed of small inorganic (PO_4^{-2}, SO_4^{-2}, NO_2^{+}, NH_3, CO_2, OH^- etc) or organic (lipids, saccharides, bases, prosthetic groups, etc) groups, ubiquitin and ubl-like domains represent the only PTM class in which the modifying moiety is a protein. This entails that, if some groups carrying out PTMs can be available as residual products of the metabolism, ubiquitin has to be synthetized as any other protein. Ubiquitination requires therefore energy for transcription and translation, making it a dispendious process. Furthermore, one ATP is consumed per ubiquitin during its conjugation. What does it make ubiquitination so important to justify such dispense of energy? Contrarily to what the name might suggest, ubiquitin is a gene present only in eukaryotes but its conservation is almost absolute from yeast to human (Zuin, Isasa, and Crosas 2014). It encodes for a 76 amino acids domain whose main function is to label proteins for proteasome mediated degradation of the target (Clague and Urbé 2010). This crucial function allows the recycling of the amino acids while maintaining the correct protein homeostasis (proteostasis) defined by the ratio between the newly synthetized proteins and the degradation of the malfunctioning ones. Proteins which are ubiquitinated are modified on lysine residues. Specifically, the ε-NH_2 group binds the ubiquitin C-terminus carboxylic group forming a so called isopeptide (Fig 2A). In order to be conjugated, ubiquitin needs the cooperation of three classes of enzymes namely called E1, E2, and E3 (Komander and Rape 2012).

Firstly, E1 enzyme uses ATP to activate ubiquitin C-terminus conjugation motif (-RGG) forming a ubiquitin-AMP derivate, which is promptly transferred onto a cysteine of the enzyme via thioester formation. The ubiquitin is then transferred as thioester moiety onto the E2 enzyme which is already competent for conjugation. The E2-ubiquitin complex finally associates with E3 ubiquitin ligase which mediates the interaction with the target proteins (Fig 2B).

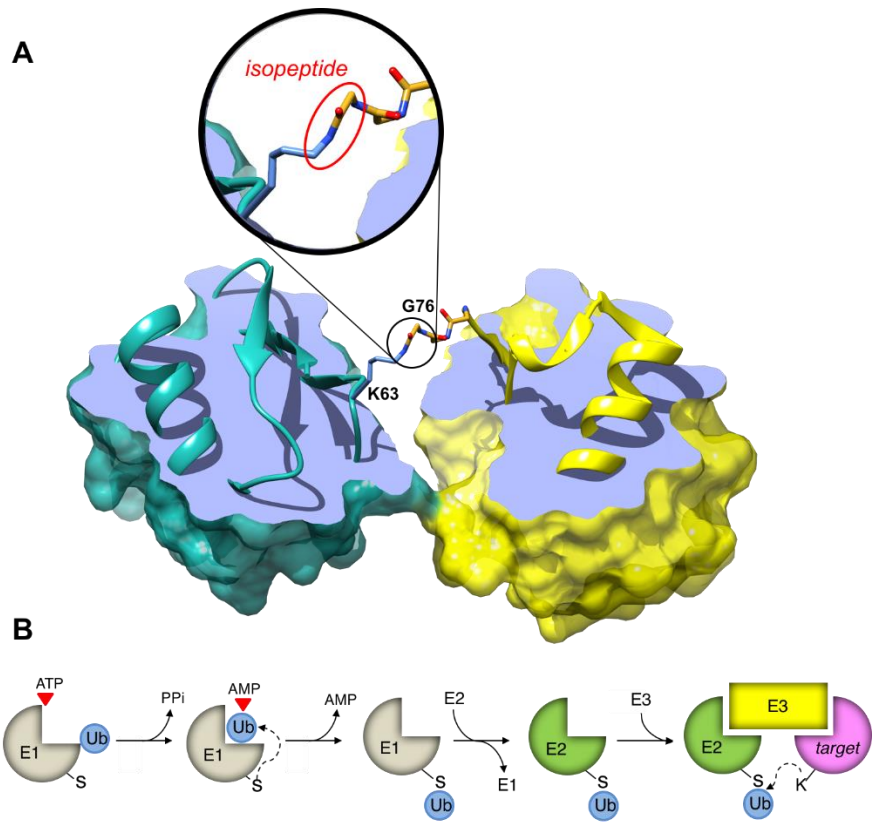


Fig 2. A) Depiction of K63-linked diubiquitin (PDB:3A9J). K63 and C-terminus G76 are shown in sticks and the isopeptide is highlighted in the zoom in circle. Donor ubq is shown in yellow, acceptor ubq in light-sea green. Isopeptide bond is circled in red. **B**) Depiction of the ubiquitination enzymatic cascade.

In humans, there are up to 1000 estimated E3 ligases (Nakayama and Nakayama 2006). Mutations associated with these enzymes are responsible for pathologies such as Parkinson and Huntington's diseases (Lipkowitz and Weissman 2011).

Target proteins can undergo mono-ubiquitination on one or different lysines as well as poly-ubiquitination on a specific residue. Meaning that, a protein baring a ubiquitin can be further ubiquitinated on one of the seven Ub-lysines (K6, K11, K27, K29, K33, K48, K63), forming a chain which can extend for several units and even include branches (Swatek and Komander 2016). Given these notions, the fate of a protein labelled with ubiquitin directly depends on: 1) Which target lysine is modified. 2) How many ubiquitin molecules are attached to it. 3) Which Ub-lysines connect the ubiquitins within the chain. For proteasomal degradation, proteins are normally poly-ubiquitinated on a single site with a K48 linked chain. This type of connectivity is by far the most frequent. Along with K48, also K11, K29 and K63 were found to promote degradation, although less frequently. As mentioned above, ubiquitinated proteins are not necessarily sentenced to death as ubiquitin regulates a plethora of other cellular functions including trafficking (K33), signalling (K29), DNA damage and innate immunity (K27), modulation of protein-protein interactions as well as protein localization (Komander and Rape 2012). In many of these non-degradative functions a major role was observed for K63 and M1 side chains which, together with mono-ubiquitination, are recognized by ubiquitin-binding domains (UBD) featured by many proteins involved in transcription, cell cycle and endocytosis (Hurley, Lee, and Prag 2006) (Hicke, Schubert, and Hill 2005).

As opposed to most of PTMs, ubiquitin is capable to display countless combinations of arrangements, each one of which imposes a different recognition mode due to the alteration of the molecular interactions. Different linkages and their combinations cause the ubiquitin monomers to interact differently within the chain, either associating tightly or by adopting more loosen conformations that consent higher degree of flexibility (Kniss et al. 2018).

In addition to ubiquitin, other domains exhibiting the same fold have been described as ubiquitin-like domains (UBLs) (Taherbhoy, Schulman, and Kaiser 2012). These domains, although share little of ubiquitin sequence, are found conjugated in a ubiquitin like fashion to many proteins involved in specific pathways such as autophagy (Atg8), nuclear-cytosolic transport and apoptosis (SUMO), and even activation of E3-ligases (NEDD8) (Kamitani et al. 1997). Lastly, ubiquitin-like domains are also synthetized as part of other proteins. In many cases, these integral UBLs play a role of regulatory domains. The function of Parkin, for example, is regulated by its integral UBL that maintains the protein in idle state via transient interactions (Sauvé et al. 2015).

1.3 Intein domains

PTMs also include proteolytic cleavage. This processing trims precursor proteins (zymogens) so to remove inhibitory portion, leading to the activation of the protein (like in the case of caspases, coagulation factors, digestive enzymes) (Khan and James 2008). The advantage of the zymogens-proteases system lies in the fact that the function of newly synthetized proteins can be put on hold until it is needed.

A similar goal is achieved by different domains, which are found translated as invading elements in other proteins sequences. When the proteins hosting such domains are translated, the presence of the invasive sequence prevents the formation of the native conformation, hence keeping the host protein in a permanent inactive state (Pavankumar 2018). The peculiarity of these invasive domains, called inteins, consists in the ability to restore the functional architecture of the host protein by exciding themselves out of the polypeptide chain, through a reaction known as protein splicing (Shao and Kent 1997). While proteases remove the inhibitory fragments by trimming terminal regions of the precursor with a single cut, inteins embody at the same time the protease and the disposable fragment as they perform a double cleavage which allows the excision of the intein itself. Concomitantly, as the reaction proceeds, the newly generated termini of the two flanks (namely N- and C-extein, respectively) are joined via the introduction of a novel peptide bond, leaving no trace of the intein insertion.

Inteins contain normally an endonuclease domain which is pivotal to the inteins life-cycle (Fig 3). Once translated, inteins use the endonuclease

domain to recognize a certain DNA sequence and perform a double strand cleavage. When the DNA-repair system is engaged, the break is resolved by homologous recombination (HR). If the donor site contains the intein, its sequence is copied into the new site. Because the endonuclease domain recognizes specific DNA sequences to cleave, once the intein has inserted, that site won't be cleavable any more while it can still be used as "donor site". Because of this, each time the intein endonuclease performs a double strand cleavage on a new site, the chance of picking a intein-containing homologous region increases until the intein propagates in every available site of the genome (Pietrokovski 2001).

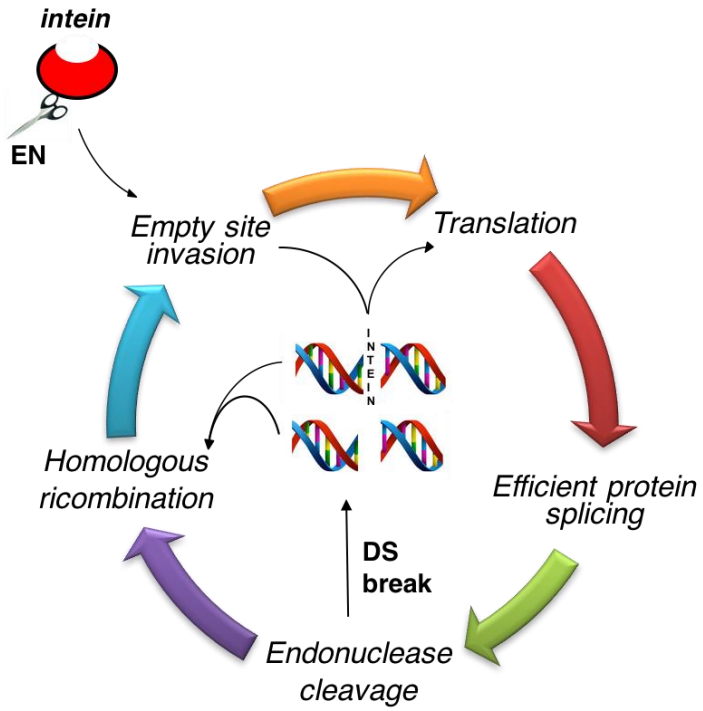


Fig 3. Graphical representation of intein invasive life-cycle. The endonuclease domain (EN) is pictured with scissors which is part of the intein fold.

For this reason, inteins have been often described as “parasitic elements”. Inteins life-cycle continues with the translation of the host protein. As mentioned before, the host protein is kept in an idle state until the intein efficiently performs protein splicing. If the host protein happens to be an essential, housekeeping protein, the lack of efficient excision will lead to cell death depriving the intein of being passed on and survive.

As opposed to this invasive life-cycle, inteins which have lost their endonuclease domains (mini-inteins) are passed on vertically and are maintained through a strong purifying selection (Soucy et al. 2014). Meaning that, whenever a beneficial symbiotic relationship between the intein and the host occurs, for that specific mini-intein the probability to be lost decreases in comparison to other inteins. An example of this co-evolution process is the existence of naturally split-inteins (Gogarten et al. 2002). In this case, the host protein (for example, DNA polymerases) is split in two genes, each one baring half intein. Only when both halves are translated, the intein reassembles and the native host protein is produced by what is known as protein trans-splicing (PTS) (Lew, Mills, and Paulus 1998). Such arrangement allows to impart a combinatorial transcriptional regulation, which wouldn’t be possible without the intein protein splicing activity.

1.3.1 Protein splicing: inside the reaction

Inteins are relatively small (~20 kDa) symmetrical domains which feature a horse-shoe shape. As they probably originate from a gene duplication,

apart for the endonuclease domain, the two halves are structurally superimposable (Fig 4A). When folded, inteins N- and C- termini are kept in close proximity to one another at the centre of the horse-shoe fold. Inteins ability to escape the precursor is based on few very conserved steps for which no cofactors are needed and no energy is consumed (Shah and Muir 2014). The first step is called N/S acyl-shift and it involves the first residue of the intein, a conserved cysteine (Fig 4B). The thiol group of this cysteine carries out a nucleophilic attack onto the backbone carbonyl group of the preceding residue, converting the connectivity between the intein and the N-extein from peptide to thioester. In this step a major role is played by a conserved T-x-x-H stretch named block B of the intein. The next step is called transesterification. During this step, a branched intermediate connecting the two exteins is formed via nucleophilic attack of the first C-extein residue (namely +1) to the thioester carbonyl. The branched intermediate still contains an ester / thioester bond, depending on the +1 residue being a serine, cysteine or threonine. At this stage, the intein is still connected with the exteins through its C-terminus. The final excision of the intein out of the precursor occurs upon the cyclization to succinimide of the conserved Asn placed before the +1 residue. As final step, called S/N acyl shift, the backbone amino group of the C-extein replaces the +1 side chain by attacking the ester / thioester carbonyl, hence restoring the peptide bond (Shah and Muir 2014). This overall recognized mechanism does not apply for rare cases in which inteins have been found lacking the +1 nucleophile residue. For these particular cases, an alternative mechanism called aminolysis was proposed (Fig 4C). According to this model, after the N/S-acyl shift the Asn cyclization

precedes the transesterification step, causing a premature C-cleavage. In absence of a +1 nucleophile, the newly generated terminal amino group of the C-extein is proposed as attacking group carrying out the transesterification step leading to the resolution of the splicing (Dassa et al. 2004).

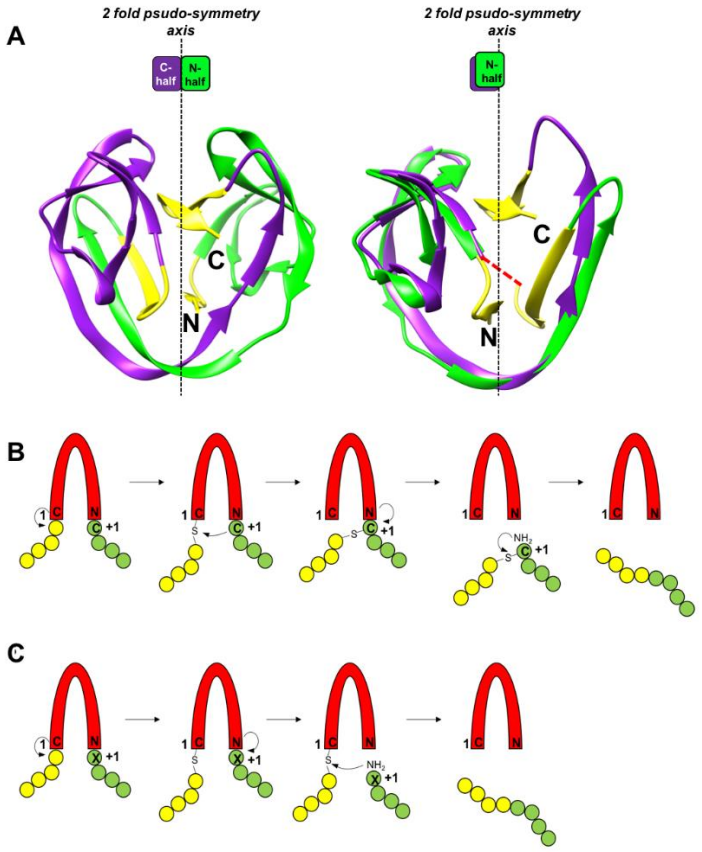


Fig 4. A) Horse-shoe like intein fold (PDB: 6QAZ). N- and C- halves are depicted in green and purple ribbon respectively. Asymmetric, non-superimposable stretches are shown in yellow. **B)** Transesterification performed by C+1 nucleophile **C)** Aminolysis proposed mechanism: the transesterification is carried out by the newly generated C-extein N-terminus. X indicates residues other than Cys/Ser/Thr. Intein is depicted in red while N- and C- exteins are shown in yellow and green, respectively.

1.3.2 Applications of inteins in biotechnology

Above all, inteins are unique because they are the only proteases which are covalently linked to their substrates. This important aspect is crucial because, upon effective protein splicing, the reaction leads to two products instead of three, being the spliced product and the intein, opposed to the protease and the two cleaved protein fragments. Another aspect of interest is that the reaction includes stable intermediates, such as the thioester bonds, whose formation can be isolated from the rest of the steps. In this way, inteins can be instructed to perform protein splicing as well as single-end cleavages. Altogether, these advantages have provided protein chemists with a new tool to improve several biochemical applications in terms of both purity and accuracy (Elleuche and Pöggeler 2010). In protein purification for example, the purified sample often requires to be cleaved away from the tag. This process is usually very delicate as the protein needs to be incubated with a specific amount of protease up to several hours, at conditions which serve to maximize the cleavage but may harm the purified protein. The cleavage requires then another step of purification, where the final protein is separated from the protease and the tag. As opposed to this routine, inteins can be used to perform an on-column cleavage, hence removing the second purification step. In this case, an N-terminus cleavage impaired intein is flanked by the protein of interest (POI) at the C-terminus and by the tag at the N-terminus (Wood et al. 1999). The protein is firstly separated by the host contaminants using the affinity for the tag until the protein is the only one remained attached to the stationary phase. The connecting peptide bond between the POI and the

intein is then cleaved upon changes in pH/temperature causing the release of the tag free POI. It's worth noticing that, if proteases can cause non-specific cleavages, inteins evade this problem because the splicing mechanism is not based on sequence recognition. Furthermore, while the quantity of protease has to be calculated based on the amount of protein, inteins are always in 1:1 ratio with the cleaving substrate. Also, the thioester formation can be controlled. Inteins preceding residues (namely -1, -2, -3) have been mutated to try to find for each intein the best chemical environment facilitating the N-S-acyl-shift. Interestingly, it was found that the C1 residue can be kept in a trapped state by di-sulphur bridge formation with another close cysteine (Callahan et al. 2011). Such bond can be displaced at leisure by introduction of a reducing agent, hence promoting the start of the reaction. Drawbacks of these strategies are mainly associated with efficiency rate (unwanted cleavage), correct fold of the intein and tolerance of POI to the splicing triggering conditions. Another application which exploits the ability of intein to perform PTS, is the segmental isotopic labelling of proteins for NMR studies. In this case, the N- and C- exteins are expressed as separate constructs each one containing half of a naturally split intein. The ^{15}N and ^{13}C labelling of one of the two constructs allows, upon successful PTS, the investigation by NMR of only a specific portion of the native protein in its natural protein environment (Liu and Cowburn 2017). The application of this method helps to overcome the problem of protein size limit in NMR, where traditional experiments generate very crowed spectra for samples above 30 kDa (Muona et al. 2010).

1.3.3 Role of exteins and inteins as PTM

The number of domains between which inteins are naturally found or artificially inserted is countless. While the fold of the inteins is conserved, the fact that extein's is not, suggests that their structural contribution within the splicing precursor is irrelevant for the accomplishment of the splicing reaction. At first glance in fact, this indication suggests that flanking domains are only accessory moieties that do not interact with inteins other than passively undergoing protein splicing. On the other hand, although they are not essential for the splicing mechanism itself, the reaction only exists because of the exteins. In fact, an intein domain alone is a completely idle element whose lack of function was agreed on by scientists, leading to the definition of inteins as "selfish element". This controversial paradox poses the question of what might be the biological role of exteins. Recent studies (Topilina et al. 2015) have revealed that exteins can participate in the splicing reaction by imposing a level of regulation which inhibits the reaction from happening in absence of specific stimuli. This type of splicing reaction falls within the definition of conditional protein splicing (CPS). The regulation control inferred by exteins turns the splicing in a non-spontaneous reaction which needs activation, given by biological stimuli for which the exteins act as sensors. Under these circumstances, the host protein can be compared to a newly released software, for the functioning of which the user has to be verified by the insertion of a "licence key". One example of extein-mediated CPS is the RadA intein from *Pyrococcus horikoshii* (Topilina et al. 2015). The exteins are two halves of a single host protein between which is interposed the intein. In

this particular case, the intein key residues responsible for the splicing are kept in a locked conformation due to polar contacts formed with the exteins residues. This conformation is stable at room temperature but upon heating (above 75° C) the exteins change their conformation and the polar contacts with the intein are lost. In the absence of inhibition, splicing spontaneously takes place and the native protein is produced. In this case, the exteins exert the function of converting the temperature shift into a biological information to which the system responds with the production of the RadA operative protein. The orthologous gene from *Thermococcus sibericus* also display the same temperature dependent CPS regulation although the triggering temperature was reported to be lower (Topilina et al. 2015). Interestingly, the different CPS activation thresholds reflect the temperature ranges in which the two organisms live, being between 70° and 102° degrees for *P. horikoshii* and between 40° and 88° degrees for *T. sibericus*. These discoveries have led to the hypothesis that the functional intein-extein partnership is the result of a co-evolution process where the intein evolves form a silent parasitic element to a controlled, protein specific, post translational modification module. Another observation in favour of this hypothesis is that inteins undergoing exteins-mediated control were found lacking the endonuclease domain. Biologically speaking, the endonuclease is lost because its function of promoting invasive gene replication is no longer needed for intein survival as the splicing regulation provides a beneficial advantage to the host.

2. BUBL: A MULTIPLE PTM PLATFORM IN EARLY EUKARYA

Few organisms belonging to a small clade of unicellular Eukaryotes called SAR (comprising *Stramneophiles, Alveolata, Rhizaria*), exclusively feature unique genes coding for multi-domain loci containing both ubiquitin and intein domains. Found in ciliates *Tetrahymena thermophila*, *Paramecium tetraurelia* and in the foraminifera *Reticulomyxa filosa*, these loci can vary in length due to the number of different domains (Fig 5B). Firstly, described by Dassa. et al, *T.thermophila* locus consists of five different ubiquitin-like (ubl), two inteins and an ADP-ribosyltransferase (ART) domain (Dassa, Yanai, and Pietrokovski 2004) (Fig 5A). These particular inteins, lacking both the endonuclease domain and the +1 nucleophile belong to a specific class of HINT domains called A-type BIL (for B*acterial* I*ntein* L*ike* domains). These are the only documented cases of bacterial intein-like elements reported outside bacteria. *P.tetraurelia* locus features only three ubl, an ART domain and one BIL, interposed between the second and third ubl module. In both *T.thermophila* and *P.tetraurelia* two divergent copies of the locus are present. Differently, *R.filosa* locus is lacking the ART domain and is constituted only of a single BIL flanked by two ubl domains. Despite this locus is not duplicated, *R.filosa* features another ubl domain which is followed by a N-terminus portion of a intein, suggesting that a PTS mechanism may be implicated. The BUBL locus (BIL-ubiquitin-Like) is considered to originate from *Paramecium* and to be conserved during the divergence of *Tetrahymena*, after which intra-genic events duplicated the number of BIL, followed by tandem duplication of the entire gene.

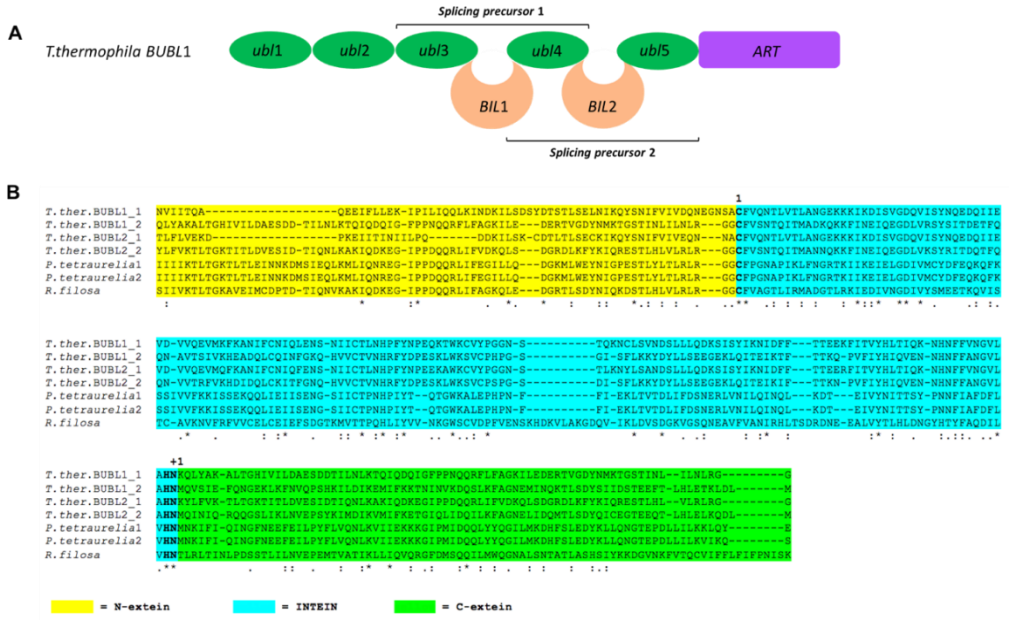


Fig 5. A) Representation of *T.thermophila* BUBL (BIL-Ubiquitin-Like) locus. BIL domains together with their direct flanking domains are highlighted as splicing precursors 1 and 2. The ADP-ribosyl-transferase domain (ART) is shown in magenta, BIL domains in salmon orange and ubl domains in green. **B)** Multiple sequence alignment performed with T-COFFEE (O'Sullivan et al. 2004) of splicing precursor systems (ubl-BIL-ubl) from *Tetrahymena thermophila*, *Paramecium tetraurelia* and *Reticulomyxa filosa* loci. N-extein are shown in yellow, C-extein in green and BIL domains in blue. Positions of catalytic residues are indicated on top of the sequences as 1 and +1.

Despite the locus has diversified in length and composition within different members of the clade, it is interesting to notice how, although not all of the BIL are preceded by ubl domains featuring the distinctive -RGG C-terminus conjugation motif (6 out of 8), all ubl ending in -RGG are followed by BIL domains (Fig 5B).

3. AIM OF THE STUDY

This study aims to elucidate the biological relevance resulting from the co-evolution of the exteins-intein symbiotic partnership within BUBL locus. In particular, the ubl4-BIL2-ubl5 splicing precursor of *T.thermophila* BUBL1 locus is investigated, giving special attention to the discovery of a novel mechanism of protein splicing for +1 nucleophile lacking BIL, which opposes the currently proposed aminolysis model (Dassa et al. 2004) and describes a new level of composite post-translational modification.

The specific aims of the studies are:

1) To investigate, using structural and bioinformatic approaches, whether the non-conjugatable ubl5 domain mediates the contact with the proteasome by interacting with its ubiquitin interaction motifs (UIM). **(I)**

2) To find, through multiple sequence alignment and comparison with structural databases, functional determinants in ubl5 sequence and structure revealing its biological role within BUBL locus. **(I)**

3) To validate using Surface Plasmon Resonance, the interaction between ubl4-BIL2-ubl5 splicing precursor and the putative partner *Tth*Ras. **(I)**

4) To study the chemistry of the splicing reaction in absence of C+1 nucleophile in presence of the native exteins and *Tth*Ras. **(II)**

5) To study of the biological role of the splicing products using computational simulations. **(II)**

6) To shed light via X-ray crystallography, electrophoresis and fluorimetry analysis on the unprecedented role of Zinc in the activation of the intein and the whole splicing precursor. **(II)**

4. MATERIALS AND METHODS

In this section, a summary of material and methods used during these studies is provided. For more detailed information, the reader is referred to the individual publications.

4.1 Molecular cloning

Synthetic genes relative to the proteins ubl5, WT precursor ubl4-BIL2-ubl5 (BUBL), the splicing hampered precursor ubl4-BIL2-ubl5-C77A/N219A (BUBL_no_sp) and *Tth*Ras were cloned as follows:

4.1.1 ubl5

Ubl5 sequence (UNIPROT_Q236S9 res 628-707) was cloned in plasmid pHYRSF53 (Kan) (Addgene # 64696) so to produce a N-terminally H_6-tagged Smt3 fusion protein. Smt3 tag was inserted to enhance the solubility of the protein and produce more starting material for the NMR structure determination.

4.1.2 BUBL and BUBL_no_sp

BUBL and BUBL_no_sp sequences (UNIPROT_Q236S9 res 410-707) were cloned in pET28b (Kan) and pET21b (Amp) as N-terminally H_6-tagged proteins, respectively.

4.1.3 *Tth*Ras

*Tth*Ras sequence (UNIPROT_I7M02) was cloned in pGEX-4-T-1 (Amp), resulting in a GST/6xHIS tagged protein. Differently from the GST, the HIS tag was not cleavable. The choice of a double tag was meant to maximize the purification efficiency.

4.2 Protein expression and purification

4.2.1 H_6-ubl5

For the expression of ^{15}N-^{13}C uniformly labelled ubl5, plasmid was transformed in *E.coli* ER2566 strain cells and grown in 2 l of M9 broth supplemented with $^{15}NH_4Cl$ and ^{13}C-glucose as sole sources of nitrogen and carbon. When OD_{600} reached ~0.6 cells were induced with a 1mM final concentration of IPTG and let grow for 4 h at 37°C. After harvesting by centrifugation, cells were resuspended in 300 mM NaCl, 50 mM sodium phosphate, pH 8 buffer and lysed by homogenization. Soluble fraction was separated from cell debris by centrifugation and the filtered supernatant was loaded onto a Ni-NTA column. The protein was purified with a linear gradient of lysis buffer containing 250 mM imidazole. The purified protein was cleaved with Upl1 and the tag was removed by a second IMAC purification. The tag-free pure ubl5 was collected in the flow-through.

4.2.2 H_6-BUBL and H_6-BUBL_no_sp

Expression of the splicing precursor required special conditions due to the intrinsic poor solubility of the multi-domain protein.

For BUBL_no_sp, transformed BL21(DE3) cells were grown in LB supplemented with 0.5 M NaCl and 1 mM betaine (LBNB) up to OD_{600} ~0.9 and induced with 1 mM IPTG, followed by 20 minutes heat-shock treatment at 47°C. Cells were then grown O/N at 20°C. The day after, cells were harvested by centrifugation and the pellet lysed by sonicated after resuspension in 20 mM Hepes, 100 mM NaCl, pH 8. After 1 h centrifugation, the filtered supernatant was loaded onto a Ni-NTA column and eluted with a gradient of buffer 20 mM Hepes, 100 mM NaCl, pH 8, 0.5 M imidazole.

For the WT precursor, in order to maximize the retrieval of all possible products of splicing, the protocol was slightly modified. When the OD_{600} reached ~0.6 the broth was added with a 0.2 M final concentration of sucrose and 3% v/v of ethanol, then induced with 1mM IPTG, heat shocked and grown as described above. Lysis and purification were carried out as above but with Tris 20 mM, NaCl 500 mM, pH 8.6, 7% glycerol buffer instead.

4.2.3 GST-*Tth*Ras-H_6

Transformed BL21(DE3) cells were grown in LB broth up to OD_{600} ~0.6 and induced with a 1mM final concentration of IPTG. After 3h at 37°C, the induced cells were collected, resuspended in 20mM Hepes, 100mM

NaCl, pH 8 and lysed by sonication. The lysate was centrifuged at 23000 g for 1h at 4°C and the soluble part was collected, filtered and loaded onto a GST-trap column. Purification was carried out by linear gradient of 20mM Hepes, 100mM NaCl, pH 8, 20mM glutathione buffer. Cleavage of the GST tag by thrombin digestion led to protein precipitation.

4.3 NMR spectroscopy

Sample for NMR analysis was prepared by dialyzing pure ubl5 against 20mM sodium phosphate, pH 6 and concentrating it up to 4 mM. A final volume of 250 µl comprising of 10% D_2O was placed in a Shigemi tube. Triple-resonance cryoprobe equipped 850 MHz AVANCE III HD and 600 MHz AVANCE III spectrometers were used for spectra acquisition which was conducted at 298K.

The following experiments were carried out for backbone assignment: [^{15}N-^{1}H]-HSQC, HNCA, HN(CO)CA, HN(CA)CO, HNCO, HNCACB, CBCA(CO)NH, while side chains ^{1}H and ^{13}C resonance assignment was based on [^{13}C-^{1}H]-HSQC, (H)CC(CO)NH, HCCH-COSY and ^{15}Nresolved[^{1}H,^{1}H)-TOCSY. Distance restrains were extrapolated from ^{13}C- and ^{15}N- edited NOESY-HSQC spectra, using mixing time of 75 and 80 ms, respectively.

4.3.1 Structure determination

Structural determination of ubl5 was performed using CYANA 3.0. NOESY cross peak assignment was used to generate two hundred initial

conformers. Additional 146 backbone angle restraints were predicted with TALOS-N. The total of 1321 restraints generated a final bundle of 20 lowest energy conformation which were further energy minimized in explicit water with AMBER14. The resulting minimized structure was analysed for structural statistics by validation tool Protein Structure Validation Suite 1.5 (PSVS). Atomic coordinated and chemical shifts were deposited in the Protein Data Bank (PDB) and in the Biological Magnetic Resonance Data Bank (BMRB) under accession codes 5N9V and 34106, respectively.

4.4 Sequence and structure alignment

Ubl5 was sequence-aligned using MAFFT (Katoh and Standley 2013) software against a collection of 78 structurally determined integral ubl domains. This library of candidates was selected by superposing ubl5 structure against ECOD (Cheng et al. 2014) and UbSRD (Harrison et al. 2016) structural databases and choosing those domains whose overall RMSD to ubl5 was below 3 Å, or which were indicated as “ubl”.
Once identified as those most similar to ubl5, members of the FERM family, were selected from a new, independent DALI server (Holm and Rosenstrom 2010) search, where ubl5 structure was used as input. After selection, FERM domains were again sequence-aligned with ubl5 (Fig 9D).

4.5 Surface Plasmon Resonance

This technique exploits the excitation of surface plasmons, quantum of oscillating electrons propagating in a parallel direction to a surface interface (i.e. solid-fluid). Each interface is characterized by a different dielectric constant. During an SPR experiment, surface plasmons are excited by incident polarized light (visible or IR). To do so, the light frequency is the same as that of the surface plasmons.

Once excided, plasmons propagate throughout the interface according to the dielectric constant. When the surface happens to adsorb some molecules, the surface plasmon wave collide with the adsorbed material which changes the dielectric constant of the interface. As result, part of the incident light is re-emitted with a certain angle and detected as RU (Resonance Unit).

Interaction between Ras and BUBL_no_sp was investigated by SPR experiments carried on a SensiQ Pioneer system. BUBL_no_sp was immobilized onto a COOH1 chip activated by injection of a 1:1 mixture of N-ethyl-N'-3-(diethylaminopropyl) carbodiimide and N-hydrosuccinimide. Immobilization was performed via amino coupling using 20mM sodium acetate at pH 4. Unreacted groups were saturated by injection of 1M ethanolamine hydrochloride. The change in resonance units (RU) was used to assess the amount of immobilized material which corresponded to 200 RU. A control cell was activated and deactivated in the same condition but without immobilizing any ligand. An equivalent internal control was introduced in the injected analytes, being GST-*Tth*Ras and GST alone. Analytes were injected at a 30 µl/min constant flow using

HSP buffer (10mM Hepes, pH 7.4, 150mM NaCl, 0.005% surfactant P20). A fast step procedure automatically diluted the analytes and injected them in serial steps. Fitting analysis were performed with SensiQ Qdat 4.0 program, using fitting for 1, 2, and 3 sites. Heterogeneous, two site curve provided the best fit.

A replica experiment was carried out in the same conditions. Analytes injections were performed at the same concentration but in HSP buffer supplemented with 50mM imidazole, in order to avoid interactions mediated by the His tag present on both BUBL_no_sp and *Tth*Ras.

The inverted experiment was carried out on a COOH1 chip activated with 1:1 mixture of N-ethyl-N'-3-(diethylaminopropyl) carbodiimide and N-hydrosuccinimide prior immobilization via amino coupling of GST and GST-*Tth*Ras. Residual groups were blocked with 1M ethanolamine hydrochloride. An activated and deactivated cell without immobilized ligand was used as internal control and used as reference for initial resonance unit estimation (25 RU). BUBL_no_sp injection was performed as fast step procedure. The sensorgrams were analysed with SensiQ Qdat 4.0 program using 1 site fitting curve.

4.6 Molecular Dynamics simulation

Atomistic molecular dynamic systems for analysis of ubl5 interaction with the Rpn10 motifs were constructed based on deposited human structures of ubiquitin in complex with proteasome ubiquitin-interacting-motifs (UIM) (PDB ID: 1YX5 and 2KDE). Structures of *T.th* UIMs were homology modelled using ITASSER (Roy, Kucukural, and Zhang 2010).

Interaction systems comprising of mutant domains or binding partners from different organisms were constructed by in silico mutagenesis and superimposition of *T.th* domains onto the deposited structures, respectively, using the CHIMERA (Pettersen et al. 2004) software.

The systems were solvated with TIP3P water molecules (Jorgensen et al. 1983) in a cubic box defined by 2nm distance between the solute and the box. Sodium and chloride ions were added to simulate a 0.1M salt concentration at electrical neutrality. All amino acids were considered in their standard protonation state that is Lys and Arg protonated, Asp and Glu deprotonated and His neutral with either delta or epsilon protonated. Verlet integrator (Verlet 1967) was used with a non-bonded interaction cut-off of 12 Å along with a 2 fs timestep. Particle Mesh Ewald (Essmann et al. 1995) method was used for long-range electrostatics. Each system was run in replica, which differed from the original run in the equilibration times (1.5 and 2.5 ns), thus resulting in a different structure for subsequent production run. Systems were simulated for 1µs at constant temperature and pressure of 310K and 1atm, respectively, for which Panariello-Rahman barostat (Bussi, Donadio, and Parrinello 2007) and Nose-Hoover thermostat (Braga and Travis 2005) were used. The protein, solvent and ions were described with CHARMM36 force field (Vanommeslaeghe et al. 2010). Analysis of the trajectories was performed using VMD (Humphrey, Dalke, and Schulten 1996). Empirical parameters, called LCF (loss of contact per frame, Eq.1) and FNB (frames not bound, Eq.2), were used to describe the binding event occurring over the simulation trajectories. %FNB was defined as the percentage of frames with less than 1 contact while LCF was defined as the summation over the number of

frames of the difference between the maximum number of contacts in the trajectory and the number of contacts in a specific frame, all divided by the number of frames:

$$\text{Eq. 1} \qquad LCF = \frac{\left[\sum_{n=1}^{nf}(C_{max} - C_{nx})\right]}{nf}$$

$$\text{Eq. 2} \qquad \%FNB = \left[\frac{(n^{\circ}\, fr_{contact<1})}{nf}\right] \times 100$$

Construction of the *Tth*Ras ubiquitinated system was carried out by preliminary docking between *Tth*Ras and ubl4 ITASSER homology model predictions (C-scores +0.65 and -0.09, respectively). Docking was performed with HADDOCK server and the isopeptide bond was simulated by setting a 2 Å unambiguous distance restraint between ubl4 G76 CO and *Tth*Ras K166 εNH2. Resulting docked structure was used as starting point for the construction of the non-standard isopeptide moiety, for which AMBER package was chosen. GROMACS based simulation was then performed using AMBER ff12SB force field. Systems were solvated with TIP3P water molecules in octahedral box, leaving 12.5Å between the protein and the box edges. Potassium and chloride ions were added at 150 mM concentration, while maintaining electrical neutrality. NVT and NPT ensembles equilibration was performed with 5 ns each, using an av-rescale thermostat at 298 K and a Berendsen pressure bath at 1 atm. 1μs production run was performed in NPT with a 2 fs time step. Particle Mesh Ewald was used for treating long range electrostatic interactions.

4.7 Liquid Chromatography-MS/MS

BUBL and the splicing-hampered mutant were separated on a 1D-gel NuPAGE 4-12% and stained with Coomassie blue. Stained bands were cut, treated first with 10mM DTT, then with 55mM iodoacetamide, and finally digested with trypsin. Peptides were analysed by Liquid Chromatography–MS/MS on an Orbitrap Fusion Tribrid mass spectrometer equipped with an Ultimate 3000 UHPLC. Peptides were desalted and then separated on a 20-cm-long silica capillary packed in-house with C18, 5 μm, 100 Å resin. The analytical separation was run for 60 min using a gradient of buffer A (5% acetonitrile and 0.1% formic acid) and buffer B (95% acetonitrile and 0.1% formic acid). Buffer B percentage increased from 5% to 30% in 35 min, then to 80% in 4 min, finally decreased to 5% concentration for a 10 min long re-equilibration step. Full scan MS data were acquired in the 350 to 1550 m/z range in the Orbitrap at 60k resolution. Data-dependent acquisition was performed using top speed mode (3 sec long maximum total cycle): the most intense precursors were selected through monoisotopic precursor selection (MIPS) filter and with charge greater than one, quadrupole-isolated and fragmented by HCD (32 collision energy). Fragment ions were analysed in the Orbitrap at 30k resolution. The AGC target value was set to 4e5 for full MS and 5e4 for MS/MS. Maximum injection times of 50 and 100 ms were used for MS1 and MS/MS, respectively. Raw data were analysed by Proteome Discoverer 2.3 using a database containing *E. coli* proteins from UniProtKB /Swiss-Prot (10577 sequences) plus BUBL sequence. Spectral matches were filtered using Percolator node, with q-values based validation and 1% FDR.

Cysteine carbamydomethylation was set as static modification while different variable modifications were considered: methionine oxidation, N-acetylation on protein terminus and the 114.043 Da mass increase on lysine residues and protein N-terminus, indicating the –RGG adduct caused by ubiquitination.

4.8 X-ray crystallography

For crystallization screening, purified BUBL_no_sp was concentrated up to ~22 mg/ml and used for testing 96 well-condition sparse matrix screens (Morpheus, $(NH_4)_2SO_4$, Index, Crystal Screen and Wizard). Spherulites appeared in Index H8 conditions (PEG 3350 15% w/v, 0.1 M Magnesium formate dihydrate) and condition optimization proceeded by hanging drop method on 24-well plates. Diffracting crystals (A & B) appeared in different wells under same optimized conditions (Magnesium formate dihydrate 0.1 M, PEG 3350 19% w/v) but with different volumes of reservoir in the well (600 and 900 ul, A and B, respectively) in order to modulate the vapor diffusion rates. Both wells were covered with 250 µl of paraffin in order to slow down the fast nucleation.

4.8.1 Data collection, analysis and structure solution

Data were acquired at 1 Å wavelength ELETTRA synchrotron XRD-2 beamline, equipped with a PILATUS 6M pixel detector. The phase problem was solved on crystal A using the Single-wavelength

Anomalous Dispersion (SAD) method. Hg derivatives were generated by soaking A-crystals in reservoir solution in the presence of 1 mM EMP (ethyl-mercuryl-phosphate) for 36 hours.
Reflection intensities were integrated and scaled using the program XDS (Kabsch et al. 2010). The two Hg sites were located by interpretation of the difference Patterson maps. The phases were calculated using the program AUTOSOL (Terwilliger et al. 2009) and model building was performed using AUTOBUILD (Terwilliger et al. 2007). The model was refined using Refmac 5 (Murshudov et al. 2011) and COOT (Emsley et al. 2010). The final structure was used as search model to solve the structure of the crystal B by using the program MOLREP (Vagin and Teplyakov 2010). Statistics of the native and heavy atom derivatives data sets are reported in Table 1 (pg. 26).

4.8.2 Fluorimetry study

Fluorescence spectra of BUBL_no_sp titrated with Zn^{2+} were collected at 25 ◦C using 1 cm path length cell, under continuous stirring. The excitation wavelength was 280 nm, and emission was recorded between 300 and 450 nm. BUBL_no_sp was equilibrated with a Chelex 100 treated buffer in order to avoid metal contamination. For titration, the protein concentration was brought to 1 μM. A stock solution of zinc chloride was prepared from atomic

absorption standard (Fluka) diluted with ultra-pure water (Fluka). The metal was added to the protein in 50 nM, 200 nM, 400 nM, 1 μM, and 2 μM increments for 4, 2, 2, 8, 8 measurements, respectively up to the final concentration of 25,4 μM.

5. RESULTS

5.1 NMR Solution structure of ubl5 (I)

Nuclear Magnetic Resonance (NMR) was used to determine the solution structure of the *T.thermophila* BUBL1 ubl5 domain. Within the entire locus, ubl5 is the only ubl domain which is not followed by a protein splicing element. Because of this, unlike ubl1/2/3 and ubl4, ubl5 is non-conjugatable as its C-terminus remains permanently connected to the ART domain, no matter the splicing events occurring within the locus. Supporting this evidence, ubl5 also lacks the –RGG distinctive C-terminus motif normally involved in ubiquitin activation (Fig 5B). These evidences indicate that ubl5 is an integral domain of an ADP-ribosyl-transferase protein. The structure determination of ubl5 was pivotal to pinpoint the differences with conjugatable domains such as ubiquitin itself. [^{15}N-^{1}H]-HSQC displayed a nice peak dispersion indicating a properly folded protein (Fig 7A).

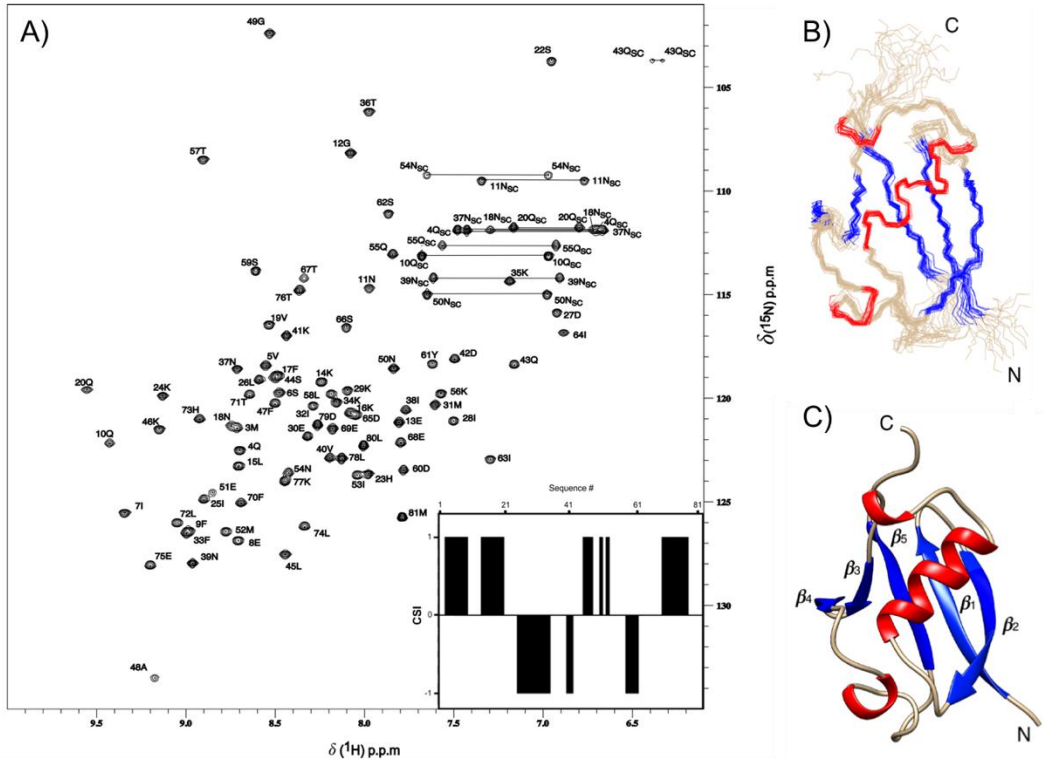


Fig 7. A) [^{15}N-^{1}H]-HSQC spectrum of ubl5. **B & C)** 20-bundle structure and lowest energy structure of ubl5, respectively. Helices are shown in red, beta-sheets in blue.

The 20 lowest energy conformers bundle shows a well conserved β–grasp fold, formed by a single helix laid on a mixed, five-stranded beta sheet with a 2-1-5-3-4 order (Fig 7B). As predicted by Chemical Shift Index analysis, additional secondary structures (helices) form between the main helix and the third beta strand as well between the fourth and fifth beta strands (Fig 7A, inset). Bundle RMSDs for secondary structure ordered regions were 0.56 and 0.92 Å for backbone and side chains, respectively (Table 1).

***T.th* BUBL1 ubl5** [a]	
Completeness of resonance assignments [b]	
Backbone (%)	99.7
Sidechain (%)	97.4
Aromatic (%)	84.9
Conformationally-restricting restraints	
Distance restraints [b]	
Total	1196
Short-range ($\|i-j\| \leq 1$)	683
Medium-range ($1 < \|i-j\| < 5$)	195
Long-range ($\|i-j\| \geq 5$)	318
Dihedral angle restraints [b]	146
Number of restraint per residue [c]	15.8
Number of long range restraint per residue [c]	3.8
Residual restraint violations	
Average number of distant restraint violations per structure [d]	
Number ≥ 0.2°	0.45 ± 0.73
Maximum (°)	0.18 ± 0.05
Average number of dihedral angle restraint violations per structure [d]	
Number ≥ 2.5°	0
Maximum (°)	0
Residual CYANA target function ($Å^2$) [b]	0.14 ± 0.005
Amber energies [d]	
Total	-73657.6 ± 2958.76
Van der Waals	12512.8 ± 570.51
Electrostatic	-97855 ± 3738.5
Model quality	
RMSD to the mean coordinate (3-65, 70-77) [c]	
Backbone (Å)	0.52 ± 0.08
Heavy atoms (Å)	0.96 ± 0.10
RMSD from ideal geometry	
Bond length (Å)	0.024 ± 0.000
Bond angles (°)	2.32 ± 0.025
MolProbity Ramachandran statistics [c]	
Most favoured	98.2 %
Allowed	1.8 %
Disallowed	0.0 %
Global quality scores (Raw / Z score) [c]	
Verify 3D	0.39 - 1.12
Prosall	0.63 - 0.08
Procheck (phi-psi)	-0.22 - 0.55
Procheck (all)	-0.21 - 1.24
Molprobity clash score	0.94 - 1.36
Model contents	
Ordered residue range [c]	3-65, 70-77
Total number of residue	81

Table 1. Structural statistics of the 20 energy-minimized conformers of ubl5. a: Structural statistics computed for the ensemble of 20 deposited structures. b: Computed with CYANA. c: Calculated using PSVS. d: Derived from AMBER.

Ubl5 structure was readily compared with ubiquitin in order to identify structural features responsible for biological function such as the binding of proteasome. Although the recognition by the proteasome would hardly be connected with protein degradation, ubl5 could serve as a localization particle leading the ART domain to modify proteasomal components via post-translational modifications.

Ubiquitin is recognized by proteasomal UIM motifs through a highly hydrophobic patch which extends over the beta sheet portion of the domain. Three key residues conserved for this patch are L8, I44 and V70. None of them are conserved in ubl5 which instead are Q8, K44 and E73, respectively.

Despite that, the composition of *T.thermophila* UIM motifs also varies from those of human so that the lack of key residues in ubl5 alone cannot demonstrate that it is unable to bind the proteasome, which may also be possible via other interfaces.

5.2 ubl5 as proteasomal localization particle (I)

In order to determine whether ubl5 could interact with UIM motifs, the two domains were studied by atomistic MD simulations in explicit solvent, in the conformation similar to ubiquitin-UIM complexes (PDB: 1YX5, 2KDE).

Ten systems were set up (Table 2). Ubl5 and ubiquitin were partnered with UIM from both *Human* and *T.thermophila* as well as with chimeras where the key residues of a UIM motif were swapped with those of the orthologous sequence (Fig 8). The same was done between ubl5 and

ubiquitin. For each complex simulated, the affinity was calculated as function of number of contacts formed over time, more precisely as *LCF* (*L*oss of *C*ontact per *F*rame) and *%FNB* (*F*rames *N*ot *B*ound).

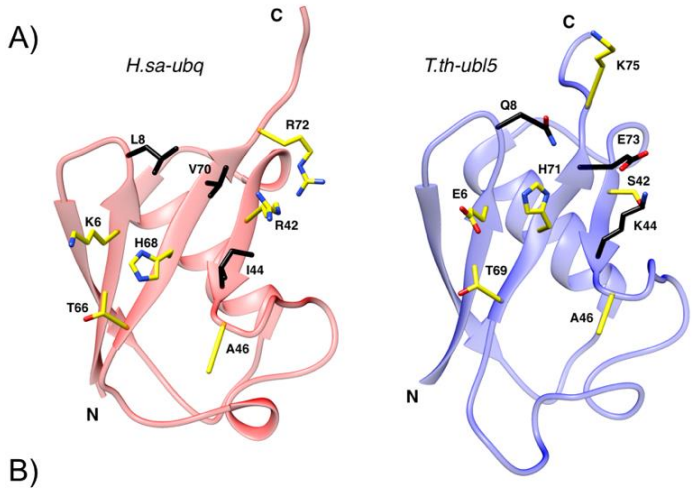
Comparative analysis of the different systems was used to assess the capacity of ubl5 to be recognized by UIM motif. As control, a reference stable interaction between human ubiquitin and UIM1 motif was described. When ubiquitin was replaced with ubl5, a destabilizing effect was introduced with two and four time increase of LCF and %FNB, respectively. Both ubiquitin and ubl5 were simulated with the reciprocal UIM partners. While the interaction of *H.sa* ubq was not affected by the UIM substitution, indicating the strong contribution of the ubiquitin key residues, the presence of *T.th* ubl5 caused an increase of the dissociation rate towards *H.sa* UIM1. This increased even further against *T.th* UIM1_a, indicating that ubl5 is unable to bind the proteasome. This was confirmed by the rescuing effect resulting from the substitution of the ubl5 key residues with those of ubiquitin.

The same set of simulations were carried out with UIM2 as well. The same discrepancy between replicas of *T.th* ubl5 / *H.sa* UIM1 is observed in *T.th* ubl5 / *H.sa* UIM2 system, although the latter (in replica a) shows a lower stability (3 to 4 times), supporting the detrimental effect due to the lack of key residues conservation on ubl5. Systems *T.th* ubl5_mut / *H.sa* UIM1 and *T.th* ubl5_mut / *H.sa* UIM2 also are comparable in term of maximum number of contacts, although the “rescuing effect” given by the humanized ubl5 was smaller for the latter. Lastly, the great resemblance between systems *H.sa* ubq / *H.sa* UIM1_mut and *H.sa* ubq / *H.sa* UIM2_mut suggests an on-off association based on conserved interactions involving

the hydrophobic patch of ubiquitin, therefore providing good control. Overall, atomistic simulations indicate that ubl5 would be unable to bind proteasomal components because of the lack of conservation of key residues forming the hydrophobic patch characteristic of the interaction cleft.

Simulation		n° LCF	$\overline{LCF}$	% FNB	$\overline{\%FNB}$	Max contacts
1	*H.sap*-ubq/*H.sap*-UIM1_(1YX5)_a	0.61	0.82	1.73	2.12	28
	H.sap-ubq/*H.sap*-UIM1_(1YX5)_b	1.04		2.56		28
2	*H.sap*-ubq/*H.sap*-UIM1-mut_a	0.81	1.74	6.23	8.71	10
	H.sap-ubq/*H.sap*-UIM1-mut_b	2.67		11.19		17
3	*H.sap*-ubq/*T.th*-UIM1_a	0.23	0.46	0.71	2.06	18
	H.sap-ubq/*T.th*-UIM1_b	0.69		3.41		16
4	*T.th*-ubl5/*H.sap*-UIM1_a	3.69	2.15	22.42	13.15	10
	T.th-ubl5/*H.sap*-UIM1_b	0.61		3.88		11
5	*T.th*-ubl5/*T.th*-UIM1_a	4.36	7.51	10.39	13.74	22
	T.th-ubl5/*T.th*-UIM1_b	10.66		17.09		31
6	*T.th*-ubl5-mut/*T.th*-UIM1_a	1.62	1.42	5.61	6.05	14
	T.th-ubl5-mut/*T.th*-UIM1_b	1.23		6.49		15
7	*T.th*-ubl5/*H.sap*-UIM2_a	15.85	8.01	75.92	39.53	20
	T.th-ubl5/*H.sap*-UIM2_b	0.18		3.14		5
8	*T.th*-ubl5/*T.th*-UIM2_a	0.01	0.04	1.72	2.71	1
	T.th-ubl5/*T.th*-UIM2_b	0.07		3.70		2
9	*T.th*-ubl5-mut/*T.th*-UIM2_a	8.86	5.24	46.67	29.16	19
	T.th-ubl5-mut/*T.th*-UIM2_b	1.63		11.66		14
10	*H.sap*-ubq/*H.sap*-UIM2-mut_a	1.88	1.70	6.49	10.91	29
	H.sap-ubq/*H.sap*-UIM2-mut_a	1.53		15.34		10

Table 2. LCF and FNB reports of the ten MD hybrid systems. Simulation replicas are named a and b for each systems. –mut flag indicates swapped residues chimeras described in Fig 9.




```
                    **       *      * **      *  *  *  *    * * ***       ***** *     * ****   *
H.sa-ubq         1 MQIFVKTLTGKTITLEVEPSDTIENVKAKIQDKEGIPPDQQRLIFAGKQLEDGRTLSDYNIQ---KESTLHLVLRLRGG
T.th-ubl5        1 MQVSIEFQNGEKLKFNVQPSHKILDIKEMIFKKTNINVKDQSLKFAGNEMINQKTLSDYSIIDSTEEFTLHLETKLDLM

T.th-ubl5-mut    1 MQVSIEFLNGEKLKFNVQPSHKILDIKEMIFKKTNINVKDQRLIFAGNEMINQKTLSDYSIIDSTEEFTLHLVTKLDLM

                    * *** ***** *   * **    **
H.sa-UIM1        1 GVDPSADPELALALRVSMEEQRQRQEE
T.th-UIM1        1 GIDPSEDPELAQAMKLSLEEEKRIQEQ

H.sa-UIM1-mut    1 GVDPSADPELAQAMRVSMEEQRQRQEE

                       ****    *   * *   *     *
H.sa-UIM2        1 LSSMTEEEQIAYAMQMSLQGAEFGQAESA
T.th-UIM2        1 FDELTEEELMNQAILASQQENPFEAQQGA

H.sa-UIM2-mut    1 LSSMTEEEQMNQAIQMSLQGAEFGQAESA
```

Fig 8. A) Structural composition of the interaction patch region for ubiquitin (left) and ubl5 (right). Core residues are shown in black, surrounding residues highlighted in yellow. B) Sequence alignment between orthologous ubq, UIM1 and UIM2 domains. Below each alignment the chimeric sequence used for MD simulations is shown.

5.3 Identification of the *Tth*Ras binding motif (I)

Based on the above indications, ubl5 was further investigated by sequence alignment analysis using two main criteria.

1) The lack of the hydrophobic patch.
2) Being an integral domain of a larger protein.

The second criterion was employed as a strategy aiming at filtering the search for similar domains through the indirect structural information

given by the ubl host protein. This is related to the fact that integral ubl domains have co-evolved with the rest of the protein in order to structurally coexist, while not compromising the overall function (Han et al. 2007). Although ubl5 is the only documented case of ubl domain integral to an ADP-ribosyl-transferase protein, looking for integral ubl domains was intended to narrow down the search to proteins in which the ubl might perform the same function of ubl5. Because structure is more conserved than a sequence, similar domains were retrieved from structural databases. Out of 78 sequences, 74 lacked of hydrophobic residues (I, L, F and C) in at least one of the three positions while only 27 belonged to integral ubl domains. Interestingly, almost half on these 27 sequences were found to belong to specific ubls, structurally part of a membrane anchoring module known as FERM featured by plasma membrane proteins (ezrin, radixin and myosin) (Frame et al. 2010) (Chishti et al. 1998) (Bosanquet et al. 2014). In particular, association with Ras GTPase was experimentally determined for SNX17 (syntaxin 17) FERM F1 subdomain (Fig 9A/B). In order to investigate whether a *Tth*Ras binding function could be hypothesized for ubl5, a multiple sequence alignment was performed between ubl5 and FERM F1 domains retrieved by structural database DALI (Fig 9D). A noticeable sequence conservation was found between ubl5 and Human SNX17 (PDB ID: 4GXB) over the second beta strand region and the helix terminus above of it (GEKLKFNV—K and GQKVLVNV—K for ubl5 and SNX17, respectively).

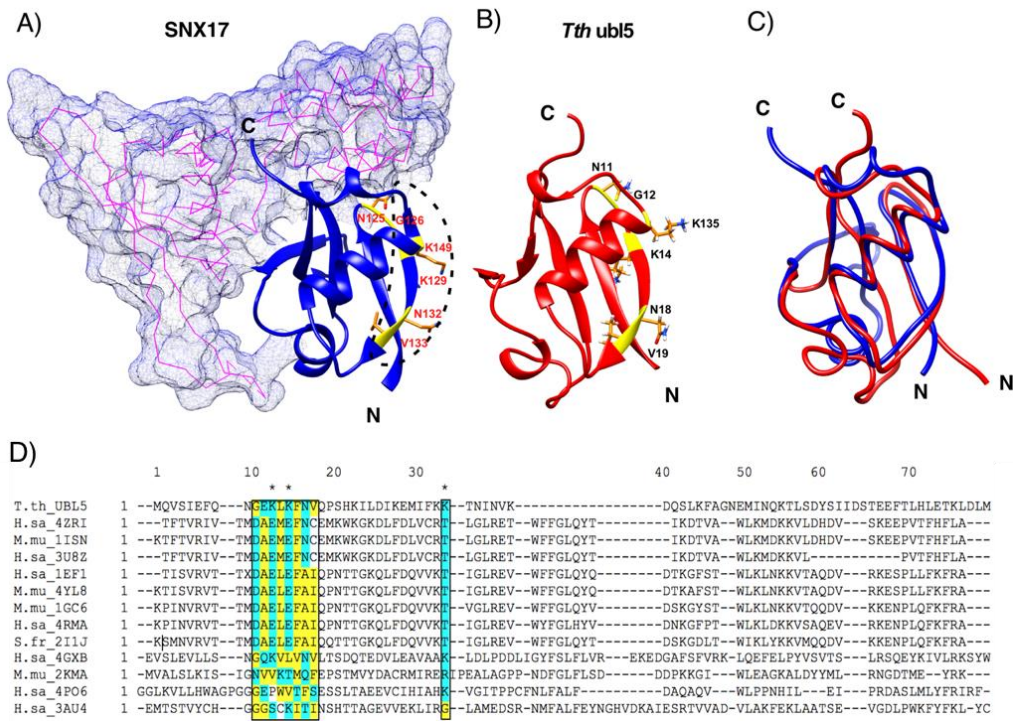


Fig 9. **A)** Structural representation of the FERM F1 module, highlighted from the SNX17 structure. Residues responsible for interaction with Ras are shown in yellow and circled by a dotted line. **B)** Structure of ubl5 with conserved residues of the Ras binding epitope highlighted in yellow. **C)** Superposition of ubl5 and SNX17 F1 domain. **D)** Sequence alignment between ubl5 and retrieved FERM domains.

Interestingly, structural superimposition of ubl5 with SNX17 F1 subdomain showed a higher conservation of the first half of the domain (that includes the Ras binding epitope) rather than the second one (Fig 9C). This is justified by the structural adaptation imposed by the surrounding protein environment of FERM over the F1 C-terminus, missing in ubl5, while the N-terminus of both domains was structurally preserved to perform the Ras-binding function.

Upon these considerations, *T.thermophila* proteome was searched for GTPases whose sequences were conserved with the H-Ras found interacting with SNX17. As documented, interaction of Ras GTPases with RBD occurs via beta strands beta strand interactions based on backbone carboxyl and amino polar groups. Nevertheless, sequence conservation was kept as search criterion. Amongst several Rab GTPases, a single Ras protein showed consistent sequence conservation of the interaction beta strand (EDSYR vs QDTYH, human and *T.thermophila*, respectively).

5.4 Experimental study of ubl5 *Tth*Ras interaction through SPR (I)

Ubl5 interaction with *Tth*Ras was established through surface plasmon resonance. The binding of *Tth*Ras to ubl5 was purposely investigated through the direct interaction of *Tth*Ras with the entire splicing hampered precursor ubl4-BIL2-ubl5 (BUBL_no_sp). This was intended to establish whether *Tth*Ras could interact with ubl5 before the splicing and, consequently, if such interaction could imply a participation of *Tth*Ras in the reaction. Main experiments were carried out by immobilizing BUBL_no_sp onto the chip and injecting GST-*Tth*Ras as analyte. Sensorgrams showed a sensible RU increment upon the injection of the analyte which could not be replicated by the injection of GST alone, indicating a tight, specific interaction between *Tth*Ras and BUBL_no_sp (Fig 10).

Interestingly, a plateau could not be observed and the best fitting curve indicated a possible double binding site. Experimental repeats were then carried out in absence and presence of imidazole (Supplementary (II), Fig

S7) to avoid possible secondary interactions caused by the His tags at *Tth*Ras C-terminus and BUBL_no_sp N-terminus, but the same trend was observed (although lessened), indicating that the secondary binding site was not caused primarily by the His-tag on either proteins.

In order to explain this peculiar interaction behaviour, we hypothesized that also ubl4 could act as RBD. Firstly, the sequence conservation of the *Tth*Ras binding epitope between ubl4 and ubl5 was looked in. Contrary to ubl5, ubl4 features on the second beta strand a distinct hydrophobic sequence (GHIVILDA—K) which would likely compromise the association with the EDSYR hydrophilic sequence of *Tth*Ras beta strand.

As ubl4 is seemingly unfit to motivate the existence of a second binding site, a reverse experiment was performed where GST-Ras was immobilized on the chip and BUBL_no_sp was used as analyte. Surprisingly, sensorgrams for this experiment were successfully fitted for a single site (Fig 11). Upon these results, it was concluded the covalent amino-coupling on chip of BUBL might have produced small populations of non-native conformers which were contributing to complex the interaction profiles. Such conformers, were absent or poorly represented when BUBL was used as analyte. In conclusion, the SPR experiments provided evidence for a specific interaction between *Tth*Ras and the splicing precursor, with a one site fitted submicromolar K_D values (500 nM ± 200 nM, *kon* 1.3 ± 0.4 10^3 $M^{-1}s^{-1}$, *koff* 6.3 ± 2.7 10^{-4} s^{-1})

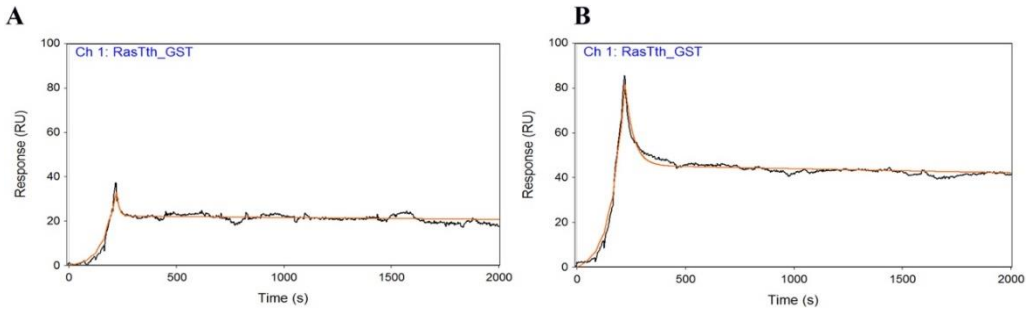


Fig 10. SPR interaction experiment. GST-*Tth*Ras in HSP buffer was injected on the sensor chip at a constant flow (30 µl/min). A FastStep procedure was used: the analyte was automatically diluted in HSP and injected by 6 serial doubling steps: 1) 0–40 s; 2) 41–80 s; 3) 81–120 s; 4) 121–160 s; 5) 161–200 s; 6) 201–220 s, where analyte concentrations were: 1) 0.0625 µM; 2) 0.125 µM; 3) 0.25 µM; 4) 0.5 µM; 5) 1 µM; 6) 2 µM (A), and 1) 0.25 µM; 2) 0.5 µM; 3) 1 µM; 4) 2 µM; 5) 4 µM; 6) 8 µM (B). The analyte is indicated in blue at the top left of the plot.

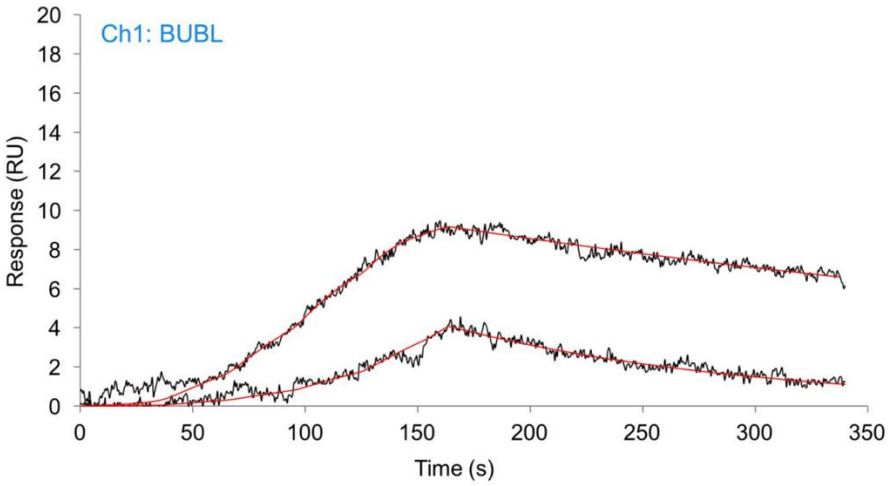


Fig 11. Control SPR experiments performed by using GST-*Tth*Ras as ligand and mutated ubl4-BIL2-ubl5 splicing precursor as analyte. A FastStep procedure was used: the analyte was automatically diluted in HSP and injected by6 serial doubling steps: At the following time points: 1) 0-30 s; 2) 31-60 s; 3) 61-90 s; 4) 91-120 s; 5) 121-150 s; 6) 151-162s, analyte concentrations were: 1) 0.0312µM;2) 0.0625 µM; 3) 0.125 µM; 4) 0.25 µM; 5) 0.5 µM; 6) 1 µM (sensorgrams A, below) or 1) 0.0937µM; 2) 0.1875µM;

3) 0.375μM; 4) 0.75μM; 5) 1.5μM; 6) 3μM (upper sensorgram B). Submicromolar KD values (KD= 500 nM ± 200 nM) and slow dissociation can be observed.

5.5 Structural characterization of the splicing products (II)

In order to understand whether the splicing reaction in absence of +1 nucleophile could take place and how, BUBL WT precursor was incubated and the splicing products were analysed by MS/MS. The first m/z match expected to be found upon the occurring of splicing corresponded to the fragment linking the N-extein C-terminus with the C-extein N-terminus. In absence of +1 nucleophile, such fragment could be formed only upon preventive C-cleavage providing for a new N-terminus amino group which directly attacks the N-junction thioester (aminolysis mechanism). Because such peptide could not be found, either protein splicing did not occur at all, or the proposed aminolysis mechanism could not explain the BUBL splicing reaction.

That is because the aminolysis requires that the C-extein remains in place, in order for the newly generated N-terminus to attack the thioester, regardless of the lack of any covalent restraint preventing its diffusion. Surprisingly, some sequences corresponding to ubl5 fragments showed a curious mass increment of 114.043 Da. Further fragmentation could assign this adduct to ubl5 K22 and K44 (Fig 12A/B). To our surprise, the 114.04 Da could be identified as a –GG (minus two H_2O lost in bond-forming condensation) sequence covalently attached to the lysines εNH_2 group. This finding provided us with the evidence that the splicing covalently connected the C-terminus of ubl4 to ubl5 lysines, and that these lysines

occupy a position within ubl5 far away from its N-terminus. Henceforth, we propose that the splicing mechanism of this +1 nucleophile lacking precursor is based on the formation of an isopeptide, where a C-extein lysine acts as nucleophile and carries out the transesterification step. Obviously, such mechanism entails a specific conformation between intein and exteins within the splicing precursor, as the C-extein nucleophile lysine falls outside the reach of the intein, which normally holds the proximal residues (…-2, -1, +1, +2…) close to one another with its fold (Fig 13).

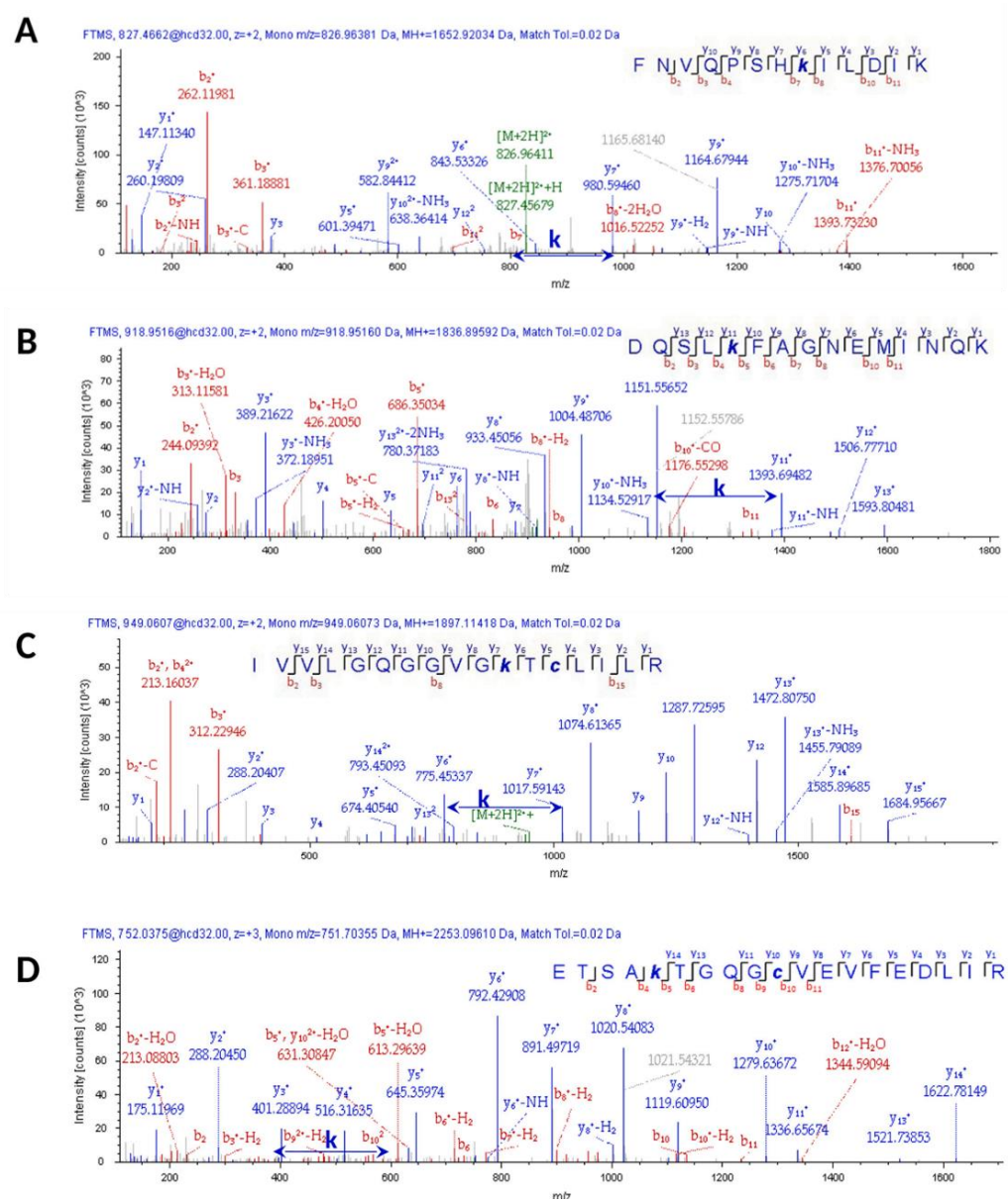


Fig 12. MS/MS spectra of A-B) the ubl5 peptides (K22 and K44) containing a -GG adduct derived from in gel trypsin digestion. C-D) *Tth*Ras ubiquitinated lysines K166 and K27. Peptide sequence is reported on top of the spectrum: **k** is for lysine modified by GG, the horizontal arrowed lines indicate b- and y-fragment ions (red and blue peaks, respectively, in the spectrum) that allow the recognition of modified lysine by a GG adduct.

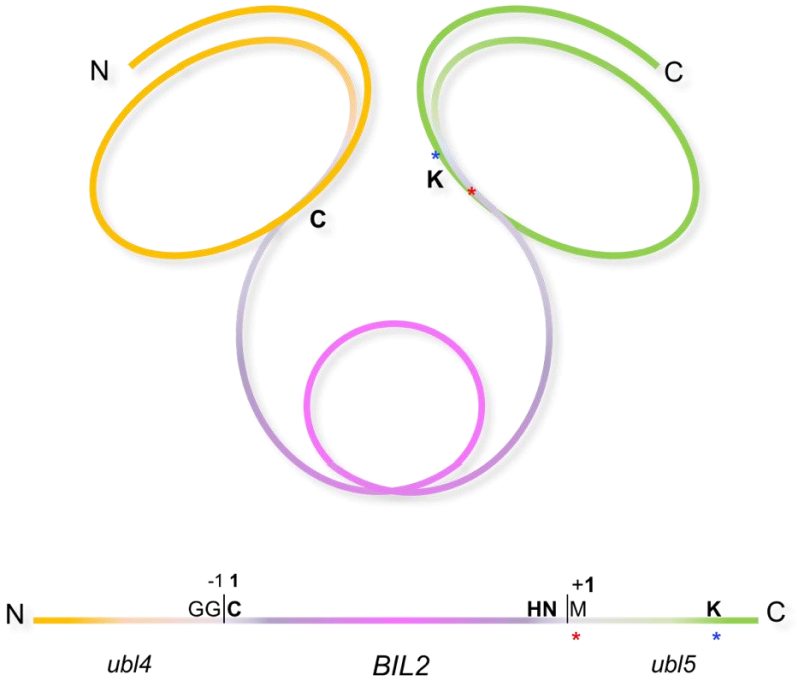


Fig 13. Cartoon depiction of the precursor tertiary conformation involved in the positioning of the catalytic residues. Protein termini are indicated as N and C. Intein is depicted in magenta shades, N- and C- exteins in orange and green shades, respectively. Intein fold alone is responsible for the proximity between residues 1 and +1 (red asterisk, normally a nucleophile). Catalytic lysine (blue asterisk) is placed far away from the 1 residue sequence-wise but structurally very close, thanks to the fold of the splicing precursor.

5.5.1 *Tth*Ras ubiquitination (II)

With the notion of *Tth*Ras interacting with the splicing precursor and that of ubl4 C-terminus being conjugated to a lysine, we investigated whether such lysine could be provided by *Tth*Ras. Therefore, we incubated the BUBL_no_sp precursor with *Tth*Ras and analysed the splicing products by MS/MS trying to identify the same -GG ubiquitination adduct on *Tth*Ras

lysines. Surprisingly, also in this case, two lysines were found carrying the di-glycine adduct. *Tth*Ras was found ubiquitinated on K166 and K27 (Fig 12C/D). Both residues are located in the so called “effector lobe” of the GTPase, responsible for the binding and the hydrolysis of the GTP and containing two regions called “switches” (Fig 14). Compared to the “allosteric lobe” the effector lobe mediates the activation of Ras via binding the GTP and inducing the interaction with the downstream effectors responsible for the signal transduction through the stretch immediately following the switch I. Hydrolysis of GTP into GDP reverses the state of the protein to inactive. Nucleotide exchange is aided by GEF (GDP to GTP) and GAP (GTP to GDP). Interestingly, K166 corresponds to the K147 of Human K-Ras, which was already known to be the target of ubiquitination (Sasaki et al. 2011).

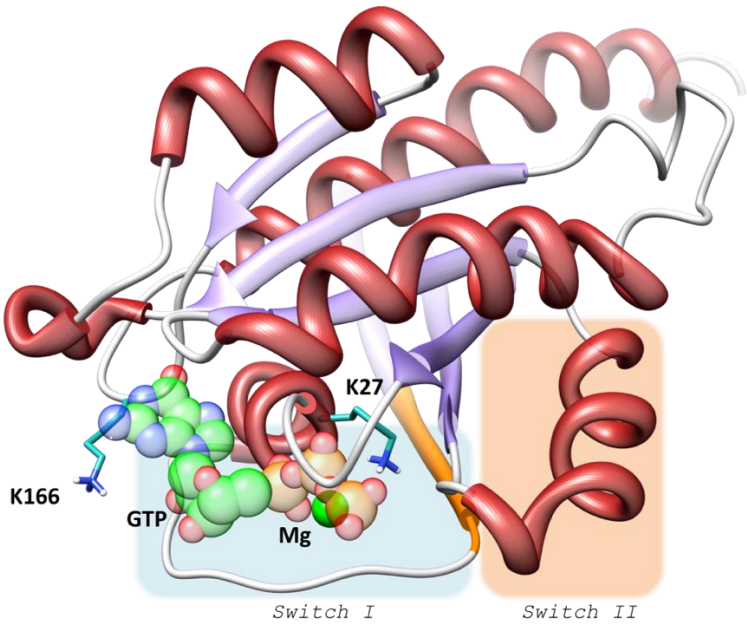


Fig 14. Homology prediction of *Tth*Ras structure. GTP and Mg are shown in spheres while switch I and II are highlighted with cyan and salmon background. Ubiquitinated lysines are shown in sticks. Effector binding stretch is shown in orange ribbons.

The possibility of this reaction occurring merely in-vitro is unlikely for several considerations. First, in absence of biological specificity, any *Tth*Ras lysine could have acted as nucleophile instead of the only two already observed to undergo ubiquitination. For the same reason, hypothetically, also any lysine of any protein from *E.coli* could have been ubiquitinated similarly. Moreover, baring ubl4 an H_6 to its N-terminus, ubiquitinated proteins from *E.coli* would have been co-purified and detected by MS.

5.5.2 Crystal structure of BIL2 (II)

Structure of BIL2 was solved at 2.3 Å resolution. Statistics are reported in Table 3.

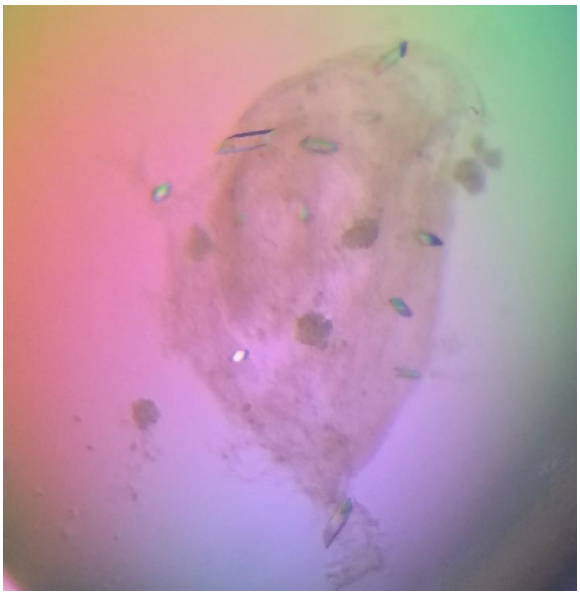


Fig 15. Crystals A of BUBL_no_sp protein.

Despite BUBL_no_sp precursor protein was used to grow crystals (Fig 15), probably due to an in-drop cleavage, only BIL2 and few N- and C-extein residues could be fit in the electron density. BIL2 crystallized in the two space groups C2 and $P2_1$ and the structures were solved by Hg based Single-wavelength anomalous dispersion (SAD) and molecular replacement (MR), respectively. The comparison of the two structures revealed no significant differences in the BIL2 overall shape which, in both

cases, nicely superposed to the only deposited structure of BIL-A domain in PDB (*C.th*BIL4, NMR structure (Aranko, Oeemig, and Iwaï 2013)) (Fig 16).

	A) BIL2 + 13/12 ext res	**B) BIL2 +10/7 ext res**
PDB code	**6TMM**	**6Y75**
Space group	C121	P1211
Cell parameters (Å)	a=133.48 b=72.76 c=70.26 90 – 97.87 – 90	a=76.82 b=67.59 c=82.42 90 – 114.57 – 90
Asymmetric unit (residues)	Tetramer (622)	Tetramer (606)
N° of bond ions	1 Ca^{+2}, 1 Hg	1 Zn^{+2}
Resolution ranges (Å)	2.4 – 50 (2.4 – 2.54)	2.3 – 50 (2.3 - 2.44)
Unique reflections	26317 (1794)	32230 (10795)
Completeness (%)	98.9 (98.2)	99.1 (99)
Redundancy	3.36	3.36
R_{merge} (%)	6 (59.3)	12.6 (61.8)
CC (1/2)	99.8 (73.3)	99 (83.1)
I/σ (I)	12.25 (1.78)	7.23 (1.87)
Resolution ranges (Å) Refinement	2.39 – 69.6 (2.39 – 2.46)	2.3 – 47.5 (2.3 – 2.35)
R_{cryst} (%)	19.7 (34.6)	25.8 (33.5)
R_{free} (%)	26.1 (42.7)	30.4 (37.6)
Rmsd (angle) (°)	1.28	0.94
Rmsd (bonds) (Å)	0.003	0.003
Wilson B-factor (Å)	58.1	32.2
Residues in core regions of the Ramchandran plot (%)	95%	98%
Residues in allowed regions of the Ramchandran plot (%)	5%	2%

Table 3. Structural statistics of BIL2 crystal structure in C2 and $P2_1$ space groups. Values in brackets refers to the last shell of refinement.

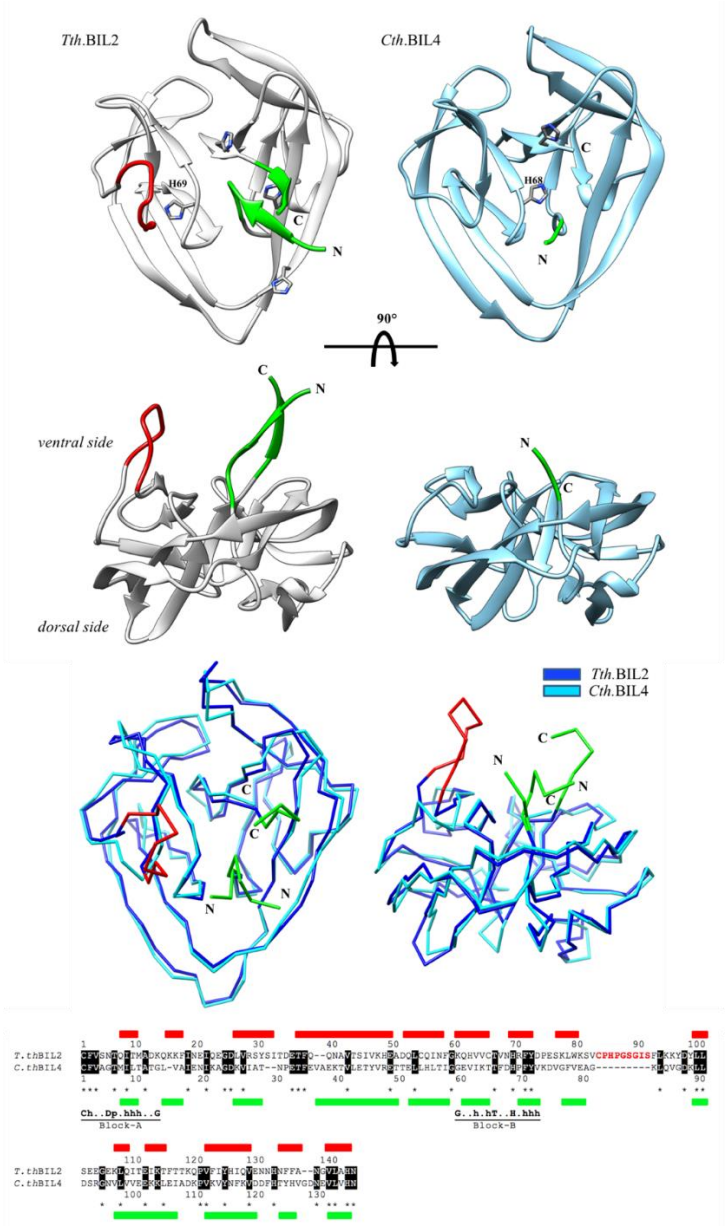


Fig 16. Structural comparison between BIL-A deposited structures (PDB: 6TMM, 2LWY). **Upper panel**: front view. **Central upper panel**: side view. **Central lower panel**: front and side view of the superimposed structures. Exteins residues are shown in green. In red is highlighted the characteristic 9 residues insertion of *Tth*BIL2. **Lower panel**: sequence alignment with secondary structure reading on top and below. Identical residues are highlighted with a star and by black background. Block A and B conserved stretches are indicated on the sequence. Green and red bars indicate beta-sheet composition.

Nonetheless, a closer view allowed to point out crucial differences regarding the side chain conformations of specific residues. In particular, the C2 structure of BIL2 was caught in an inactive state where the block-B H69, essential residue responsible for the stabilization of the N/S-acyl shift intermediate, is placed away from the N-terminus splicing junction. As no similar conformation could be found in any previous intein structures, we concluded that BIL2 intein is inactive by default (Fig 17A). Confirmation of this model was provided by the comparison with the $P2_1$ structure, in which H69 was found in the common, catalytic conformation, with the side chain pointing towards the intein N-terminus (Fig 17B).

In order to discriminate the factors responsible for this active conformation, crystal packing was scrutinized so to identify chain-chain contacts involving the H69 region though none appeared existing. Next, electron density was analysed for identification of ligand molecules such as reservoir components or buffers, which may have participated in the structure stabilization. To our surprise, a large missing electron density was found on top of histidines 125 and 48 both arranged in a coordinating-fashion of what it seemed a metal. As *fo-fc* density signal could be observed until ~12 sigma, a heavy-atom of around 30 of atomic number was hypothesized, though spectroscopic analysis could not be conclusive about its identity. In support of the presence of a metal atom, the anomalous difference map derived from 1Å data also showed a clear electron density (paper2, Fig S4). Based on the fact that residual Ni from the IMAC purification had been removed with dialysis during sample preparation and that it had been reported that lab plastic releases Zn^{+2} ions, we concluded that the density was Zn^{+2} from the crystallization plate.

Furthermore, BIL2 is not the first intein to be crystallized with a bound Zn^{+2} ion which was not originally included in the sample or reservoir composition (Nichols et al. 2003) (Mills and Paulus 2001).

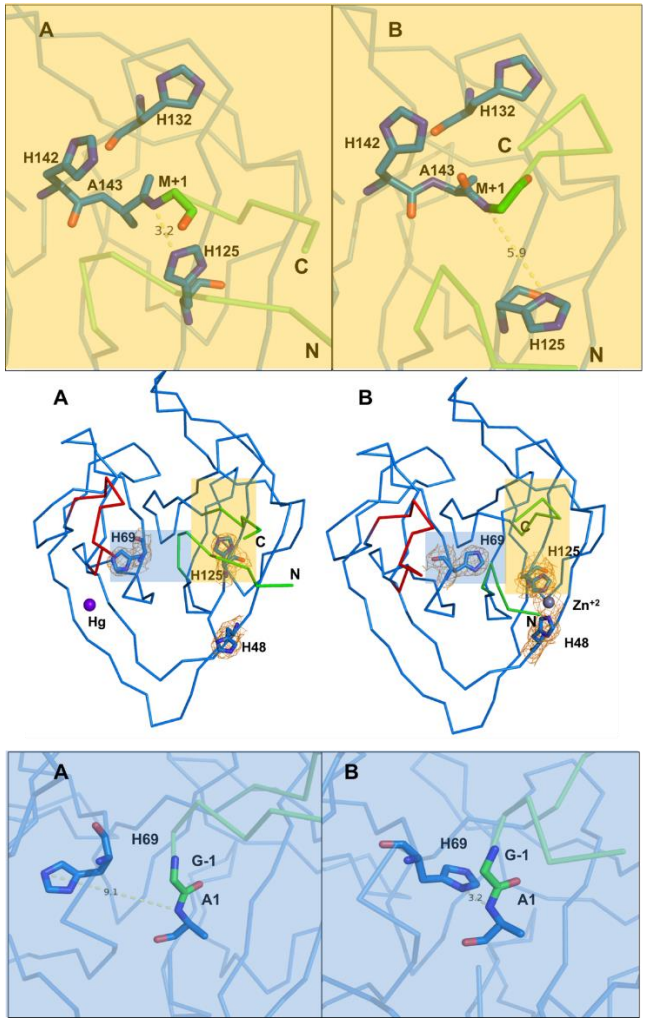


Fig 17. Comparison of BIL2 crystallographic structures. **A)** C2 space group Zn^{+2}-free. **B)** P21 space group, Zn^{+2}-bound. Middle panel: the CPHPGSGIS insertion sequence is shown in red while exteins residues are shown in green. Residues affected by Zn^{+2} binding (H69-H125-H48) are represented with sticks and contoured by electron density map at 1 sigma. Zooms of the N- and C-splicing junction are shown in light blue and gold, respectively.

5.5.3 Structural insights of Zn^{+2} dependent BIL2 activation (II)

Differently from other intein structures deposited on PDB, the Zn^{+2} binding site of BIL2 is far away from the splicing junctions. In particular, the ion is bound in a square-pyramidal penta-coordinated fashion by residues H125, H48, N-extein N-5 NH and CO backbone groups, and either E23 form a crystal symmetric chain or a water molecule (Fig 18).

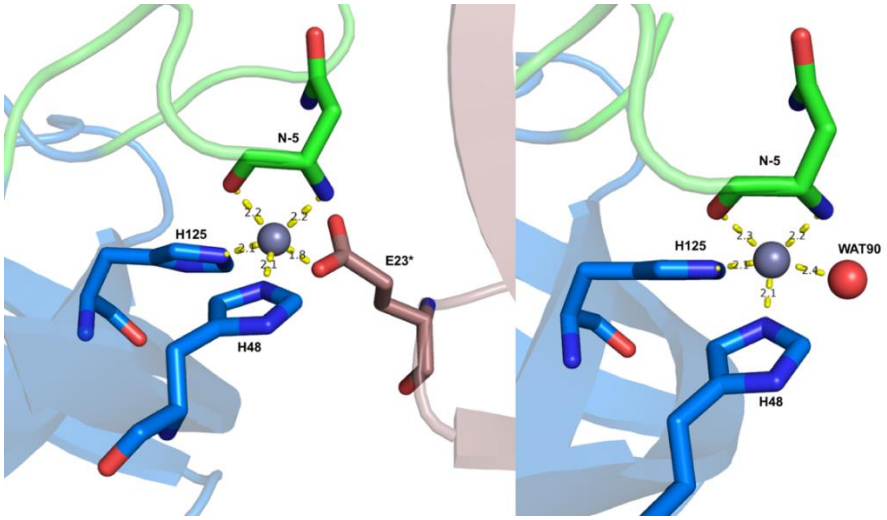


Fig 18. Zn^{+2} coordination via crystal contacts (left panel) and water mediated (right panel). Both cases reflect a square pyramidal penta-coordination, with the ion at the centre of the square base. Residues are indicated by one-letter code and number. Distances of the ion from coordinating group are shown. Asterisk indicate residue from symmetrical chain. BIL is depicted in marine blue, exteins in green.

Comparison with the apo structure revealed the hydrogen bond network connecting the Zn^{+2} binding site to the activation of H69 (Fig 19). First, upon Zn^{+2} binding, the flanking extein residues dramatically rearrange,

meaning the N-extein switches from forming a beta sheet interaction with the C-extein to being dragged away from it and moved towards the H125 and H48, which adopt a closed conformation.

In both cases, stable hydrogen bonds occurred between L-4 CO and G-1 NH and between T66 γCO and A1 NH (Fig 19). Contrarily, two alternative conformation of N68 could be seen for the apo structure where only one is present in the presence of Zn^{+2}. This is due to additional h-bond coordination from L-4 CO. Altogether, this stabilized N68 conformer seems to facilitate the new arrangement of the following residue (H69) so that its backbone NH forms and hydrogen bond with T66 CO and the side chain flips towards the G-1/A1 peptide bond (Fig 19).

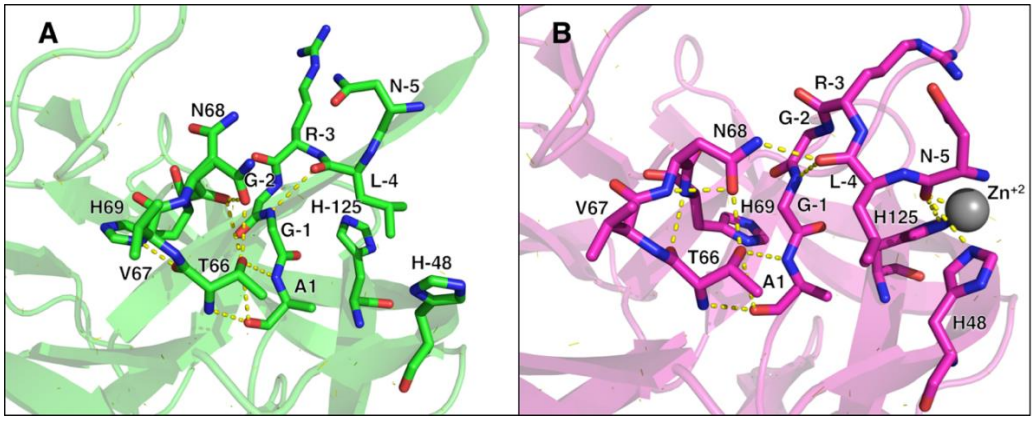


Fig 19. Highlights of the hydrogen bond network in absence **(A)** and presence **(B)** of Zn^{+2}. Hydrogen bonds are represented with yellow dotted lines and residues are indicated with one-letter code followed by numbering.

Besides the activating mechanism through which H69 adopts its catalytic conformation, Zn^{+2} may have an additional role in delaying the C-cleavage reaction. Specifically, upon Zn^{+2} binding, H125 is seen moving away from

the C-terminus scissile bond, while its position is favourable for C-terminus scissile bond stabilization (3.2 Å from the M+1 NH) in the apo form. In this way and under no control, H125 may trigger premature C-cleavage preventing catalytic ubl5 lysines to perform transesterification (Fig 17A, upper panel).

5.6 In vitro Zn^{2+} dependent BIL2 activation (II)

Assuming that the precursor may not have a defined preferred conformation, Zn^{+2}-induced activation of protein splicing cannot rely on the H69 side chain flip alone, but needs to impart a wider effect by which the exteins are arranged in a conformation where the distant, catalytic lysine is brought close to the thioester. Such large motion within the precursor is already hinted from the structure where exteins are seen to undergo larger conformational changes upon Zn^{+2} binding than the intein. In order to validate the effect of Zn^{+2} on the splicing reaction in its entirety, the WT precursor was incubated with different concentration of Zn^{+2} acetate and, after analysis of SDS-PAGE, the relative intensity of the products bands was quantified. In particular, ratios were calculated by dividing the intensity of the splicing precursor band by that of BIL2 as well as of the linked exteins bands (ratios are called P/I and P/Is, respectively) (Fig 20).

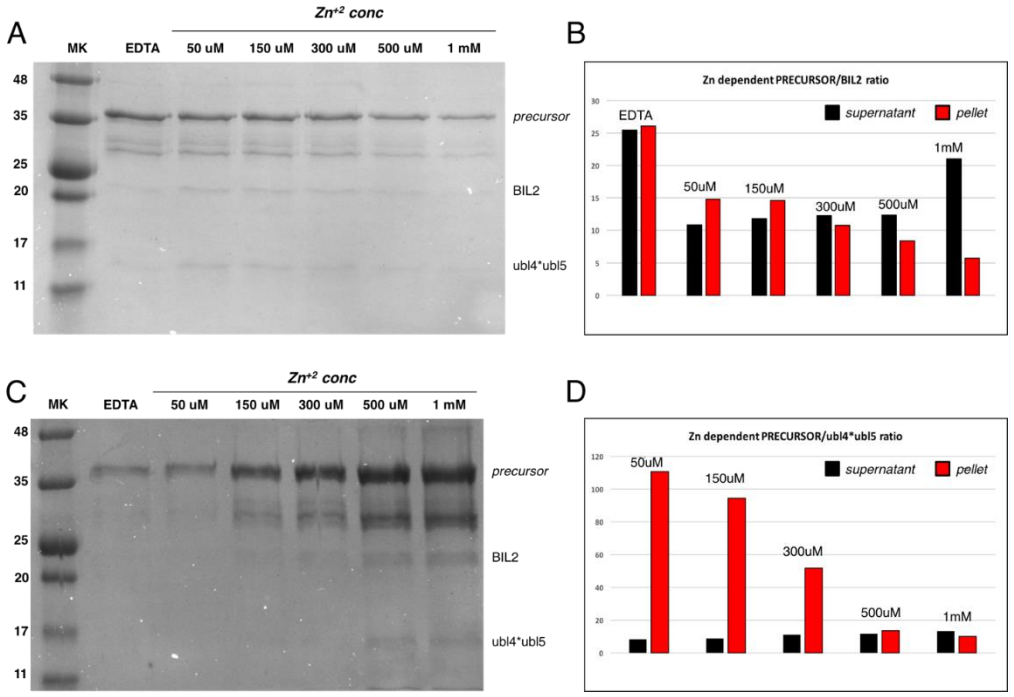


Fig 20. Incubation of BUBL-WT with Zn^{+2} concentrations. **A)** supernatant fractions after 48 h incubation. **B)** Zn^{+2}-dependent ratio between splicing precursor and BIL2 (P/I) in the two fractions. **C)** pellet fractions after 48 h incubation. **D)** Zn^{+2}-dependent ratio between splicing precursor and splicing product (P/Is) in the two fractions. Ratios were calculated with ImageJ and the graphs elaborated with Excel.

The effect of Zn^{+2} binding on the precursor appeared evident as increasing level of precipitation were observed at increasing Zn^{+2} concentrations. Due to this precipitation, samples were centrifuged and SDS-PAGE analysis was carried out on both soluble and precipitated fractions.

WT BUBL incubation resulted in the production of five main bands. While in the soluble fraction, the P/I increased with Zn^{+2} concentration indicating that the relative amount of free-BIL2 decreased, in the precipitated

fraction, the opposite trend was observed. In fact, the P/I ratio decreased as the free-BIL2 kept accumulating. The fact that soluble and precipitated fractions showed opposite trends, supports that precipitation occurred upon Zn^{+2} binding.

When the P/Is ratio were analysed, the same opposite trend could be found between the two different fractions, although with very different slopes. This could be argued with the fact that P/Is only accounts for efficient protein splicing while P/I it's representative of either splicing, N- and C-cleavages. Overall, the amount of splicing product rapidly increases in the precipitate fraction while it slowly decreases in the soluble one. This evidence suggests that while Zn^{+2} can directly control the N/S acyl-shift by interacting with the intein residues, it also can regulate the C-cleavage in a fashion that probably involves extein portion which could not be observed in the structure. Overall, the Zn^{+2} regulation of the splicing reaction works on different levels: first, Zn^{+2} binds BIL2 and prevents the C-cleavage by coordinating H125 while it activates H69. As progressively Zn^{+2} is bound to other sites, it concurs to alter the precursor conformation in order to place the catalytic lysine close to the N-terminus splicing junction. The same conformational change, possibly promotes an interaction between the C-extein and BIL2 which leads to the conclusive C-terminus cleavage.

5.7 Fluorimetry study on Zn+2 affinity and stoichiometry (II)

In this experiment, in order to isolate the Zn^{+2} binding from the induced downstream splicing events, the BUBL_no_sp mutant was used. Normalized intensity over Zn^{+2} concentration clearly showed three plateaus corresponding to an equivalent number of binding sites with decreasing affinities. Fitting of the higher affinity site allowed to calculate a K_D of 1.2 ± 0.2 μM.

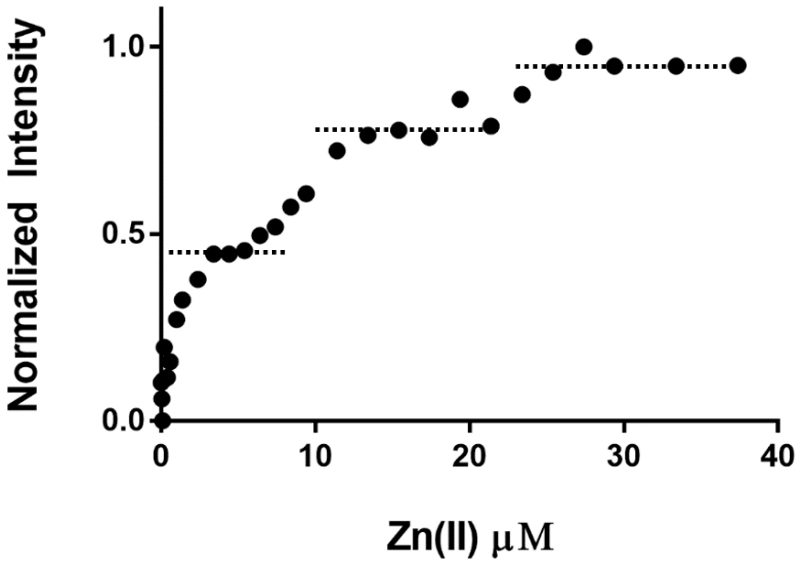


Fig 21. Complete titration of BUBL_no_sp with Zn acetate. The three plateaus are highlighted by dotted lines.

The discovery of a 3:1 (Zn^{+2} to protein) stoichiometry on the whole precursor, jointly with the 1:1 observed from crystallographic data for

BIL2 alone, allows to conclude that Zn^{+2} binds the exteins on at least two different sites. Although the question of on which domain resides the higher affinity site could not be answered, it is remarkable to notice how Zn^{+2} progressively binds the precursor at different sites in a specific order. This, together with the electrophoresis profiles, indicates how each of the sites introduces a modification of the precursor structural arrangement resulting in precipitation and splicing.

5.8 Biological implication of intein mediated ubiquitination (II)

While the E3 ligase responsible for K-Ras ubiquitination on K147 in Human remains unknown, the conserved lysine is ubiquitinated in *T.thermophila* by a protein splicing platform which is able to lure the ubiquitination target before transferring the ubl moiety. Of the two *Tth*Ras lysines subjected to ubiquitination, K27 resides within the nucleotide binding pocket and therefore, its ubiquitination would completely abolish the possibility to bind the nucleotide. On the other hand, K166 is located at the marginal side of the biding pocked where it interacts with the guanosine side of the GTP. In order to assess how the ubiquitination of K166 modifies the state of *Tth*Ras, the conjugated system ubl4-K166-*Tth*Ras was investigated by molecular dynamics with GTP and GDP and compared against unconjugated GTP and GDP-bound systems.

After 1μs long simulation in explicit solvent, conjugated systems displayed lower RMSF per residue in comparison to the unconjugated systems, proving the stabilizing effect of ubl4 (Fig 22). Such stabilization seemed to be also due, in part, to the bound nucleotide. In particular, GTP seemed

to stabilize more the complex than GDP. The reason for this additional contribution lies in the effect of the γ-phosphate hindrance at the ubl4-*Tth*Ras interface. In fact, while in the GDP system the absence of the γ-phosphate allows to accommodate the switch I Y73 aromatic ring within the nucleotide binding cleft, in the GTP system the Y73 side chain is pushed outside, towards the solvent. Although, in the absence of ubl4 this different conformation does not influence the RMSF values throughout the protein, in conjugated system it provides the anchor through which the ubl4 tightens the interaction with *Tth*Ras, leading to a much greater stabilization. A closer view reveals that, in presence of ubl4, Y43 is involved in a very stable cation-π interaction with the guanidinium group of ubl4 R42 (Fig 23). Analysis of centroid-centroid distance and plane-plane angle between the R42 guanidinium and the Y43 aromatic ring revealed that the interaction remains stable during the whole duration of the simulation. Confirmation of the stabilizing role of this interaction can be drawn from the GDP conjugated system, where the lack of interaction causes the ubl4 moiety to oscillate more with respect to *Tth*Ras, within the limits of the covalent isopeptide constrain. Overall, based on the scientific knowledge about small GTPases, the ubiquitination of *Tth*Ras K166 clearly causes an activation of the protein with respect to its ability to interact with downstream effectors. In fact, in presence of GTP, ubl4 prevents its hydrolysis by sequestering the Y43 and therefore, locking *Tth*Ras in a perpetual active state, probably until its deubiquitination.

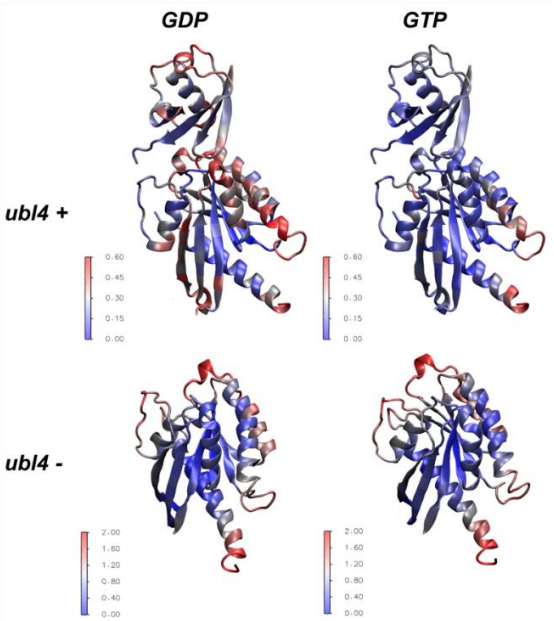


Fig 22. Structural depiction of C-alpha RMSF values for conjugated and unconjugated systems, bound to GTP and GDP, with respect to the simulation starting structure. RMSF values are indicated by a colour-based scale (0.0 to 2.0 for unconjugated systems, 0.0 to 0.6 for conjugated systems).

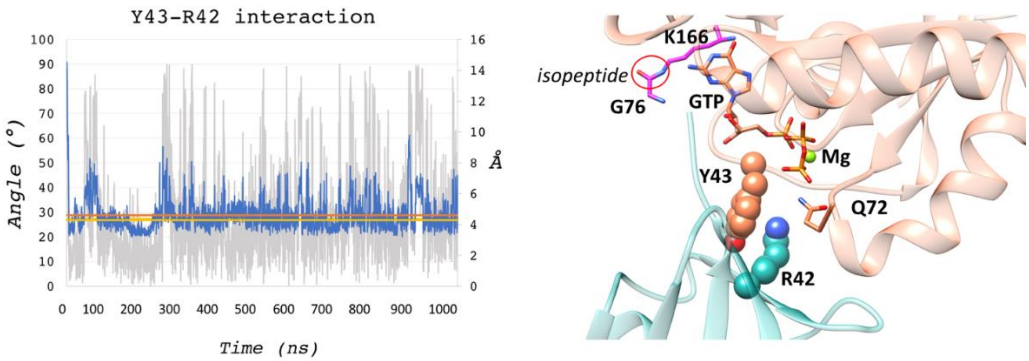


Fig 23. Left panel: Plane-plane angle (grey) and centroids distance (blue) between *Tth*Ras Y43 aromatic ring and ubl4 R42 guanidinium group over simulation time. Mean values are shown in yellow and orange, respectively. **Right panel**: Structural representation of the interaction. *Tth*Ras Y43 and ubl4 R42 are shown in coral and light sea green spheres, respectively. GTP is shown in sticks as Q72 and the magenta G76 K166 residues linked by the red-circled isopeptide.

6. CONCLUSIONS

Inteins were initially described as parasitic elements, which spread undisturbed across the genome by passing undetected the host organism defences. This means that, through efficient protein splicing, inteins have managed to prosper within that very narrow blind spot in the cellular control system which is the ability to leave the host protein unscathed and functional. Because of that, inteins have always used the synthetic and replicative machinery of the host organism for free, without even the need to highjack it, as it occurs in most of virus-host relationships. In this case, a perfect tolerance was developed towards inteins, which highlights an interesting evolutionary perspective, according to which: "molecular conservation does not imply biological function, as well as the lack of biological function does not justify, per se, molecular loss." The condition where an element is maintained without providing advantage to the host is represented in Fig 24 by the white region. Alterations of this state of equilibrium can result in two scenarios, depending on whether the intein proves to be advantageous or detrimental to the host. In the latter case, if the host fitness is compromised so is that of the intein, therefore causing its loss. This is the reason why it is correct to presume that, at a certain time, all observable inteins have adapted themselves either to benefit the host or, at least, to not harm it. Functions benefitting the host include C-terminus lipidation (Sonic Hedgehog) (Perler 1998), protein synthesis regulation by dual-transcriptional activation (PTS) (Ciragan et al. 2016) or temperature shifts (*Pho*RadA) (Topilina et al. 2015).

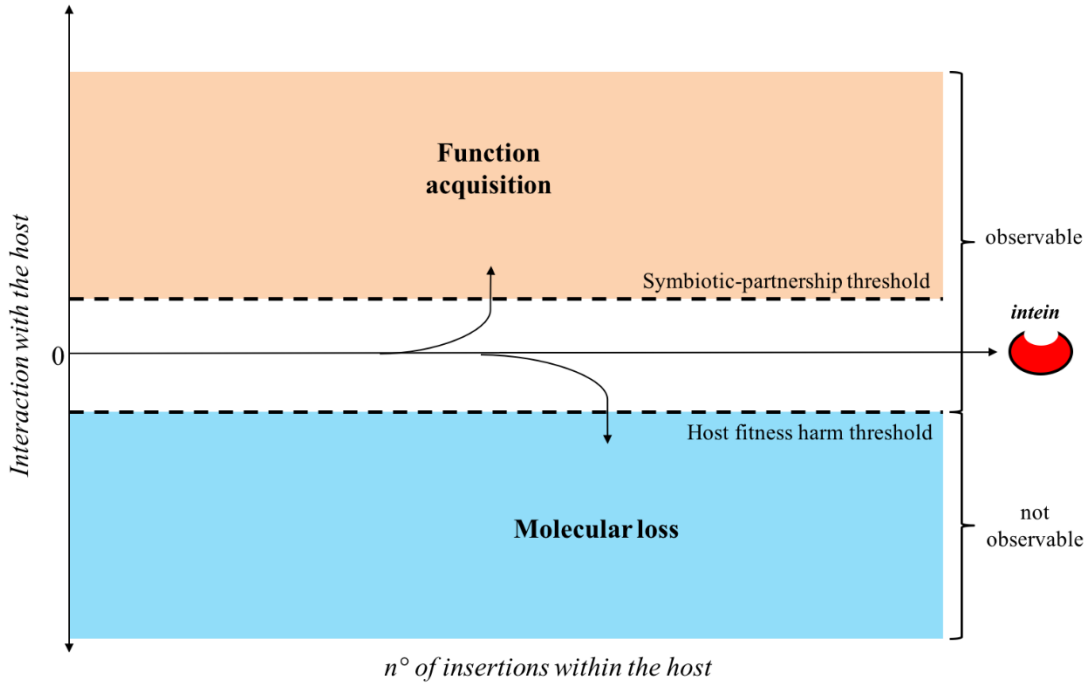


Fig 24. Graphical depiction of the intein evolutionary landscape as a function of both n° of insertions and degree of interaction with the host.

In this work, the symbiotic partnership between the host and the intein is described to occur at a new level, where the protein splicing does not simply remove the obstacle which keeps the host protein in an idle state, but is responsible for the structural editing which confers biological function to it. By changing the connectivity between ubl4 and ubl5, BIL2 bans a specific pool of conformations while fans out specific different ones. The way two ubiquitin-like domains interact with each other according to their covalent connectivity, determines the overall surface composition of the chain and its mobility, which sets the target of the interaction as well as the degree of the affinity. Thus, in this unprecedented

case, the host protein prior the intein insertion is not equivalent to the protein after the intein excision.

Interestingly, such reaction is still aided by a chemical input, being the presence of zinc ions, which stabilize the active form of BIL2 while favouring the orientation of the catalytic lysine. This indicates that a high level of control was developed by the synergistic coevolution between the intein and the exteins, which could not be foreseen by the splicing precursor sequence, outlining how, for multi-domain proteins the structural interdependence between domains plays a greater role than the individual conserved function. That is furthermore substantiated by the ability of the splicing precursor architecture to bind *Tth*Ras GTPase with high affinity and in such orientation to be itself ubiquitinated on a conserved lysine found seemingly modified also in complex Eukarya (Human). The biological consequences caused by the ubiquitination of *Tth*Ras appear to be consistent with the stabilization of the active form found for ub-K-RAS. This, adds a further level of complexity, making BUBL a real PTS platform of which much remains unknown yet. In fact, BIL1 and BIL2 could both act in absence of *Tth*Ras, producing an heteromeric, branched ubiquitination of the integral ubl5 domain (ubl1-ubl2-ubl3*ubl4*ubl5-ART), as well as the ADP-ribosylation of the lured *Tth*Ras.

In conclusion, in this work, it is shown how a viral element (intein) was functionalized, by exploiting the chemistry of its survival cycle, to carry out essential biological functions, normally performed only by specialized enzymes. Such event, originated by the serendipitous insertion of the intein, represents a unique and remarkable example of molecular evolution.

References:

Aebersold, Ruedi, Jeffrey N. Agar, I. Jonathan Amster, Mark S. Baker, Carolyn R. Bertozzi, Emily S. Boja, Catherine E. Costello, et al. 2018. "How Many Human Proteoforms Are There?" *Nature Chemical Biology*. https://doi.org/10.1038/nchembio.2576.

Ambrogelly, Alexandre, Sotiria Palioura, and Dieter Söll. 2007. "Natural Expansion of the Genetic Code." *Nature Chemical Biology*. https://doi.org/10.1038/nchembio847.

Aranko, A. Sesilja, Jesper S. Oeemig, and Hideo Iwaï. 2013. "Structural Basis for Protein Trans-Splicing by a Bacterial Intein-like Domain - Protein Ligation without Nucleophilic Side Chains." *FEBS Journal*. https://doi.org/10.1111/febs.12307.

Bosanquet, David C., Lin Ye, Keith G. Harding, and Wen G. Jiang. 2014. "FERM Family Proteins and Their Importance in Cellular Movements and Wound Healing (Review)." *International Journal of Molecular Medicine*. https://doi.org/10.3892/ijmm.2014.1775.

Braga, Carlos, and Karl P. Travis. 2005. "A Configurational Temperature Nosé-Hoover Thermostat." *Journal of Chemical Physics*. https://doi.org/10.1063/1.2013227.

Bussi, Giovanni, Davide Donadio, and Michele Parrinello. 2007. "Canonical Sampling through Velocity Rescaling." *Journal of Chemical Physics*. https://doi.org/10.1063/1.2408420.

Callahan, Brian P., Natalya I. Topilina, Matthew J. Stanger, Patrick Van Roey, and Marlene Belfort. 2011. "Structure of Catalytically Competent Intein Caught in a Redox Trap with Functional and

Evolutionary Implications." *Nature Structural and Molecular Biology*. https://doi.org/10.1038/nsmb.2041.

Cheng, Hua, R. Dustin Schaeffer, Yuxing Liao, Lisa N. Kinch, Jimin Pei, Shuoyong Shi, Bong-Hyun Kim, and Nick V. Grishin. 2014. "ECOD: An Evolutionary Classification of Protein Domains." Edited by Arne Elofsson. *PLoS Computational Biology*. https://doi.org/10.1371/journal.pcbi.1003926.

Chishti, A H, A C Kim, S M Marfatia, M Lutchman, M Hanspal, H Jindal, S C Liu, et al. 1998. "The FERM Domain: A Unique Module Involved in the Linkage of Cytoplasmic Proteins to the Membrane." *Trends in Biochemical Sciences*. https://doi.org/10.1016/S0968-0004(98)01237-7.

Ciragan, Annika, A Sesilja Aranko, Igor Tascon, and Hideo Iwaï. 2016. "Salt-Inducible Protein Splicing in Cis and Trans by Inteins from Extremely Halophilic Archaea as a Novel Protein-Engineering Tool." *Journal of Molecular Biology*. https://doi.org/10.1016/j.jmb.2016.10.006.

Clague, Michael J., and Sylvie Urbé. 2010. "Ubiquitin: Same Molecule, Different Degradation Pathways." *Cell*. Cell Press. https://doi.org/10.1016/j.cell.2010.11.012.

Dassa, Bareket, Haim Haviv, Gil Amitai, and Shmuel Pietrokovski. 2004. "Protein Splicing and Auto-Cleavage of Bacterial Intein-like Domains Lacking a C-Flanking Nucleophilic Residue." *Journal of Biological Chemistry*. https://doi.org/10.1074/jbc.M404562200.

Dassa, Bareket, Itai Yanai, and Shmuel Pietrokovski. 2004. "New Type of Polyubiquitin-like Genes with Intein-like Autoprocessing

Domains.” *Trends in Genetics*. https://doi.org/10.1016/j.tig.2004.08.010.

Elleuche, Skander, and Stefanie Pöggeler. 2010. “Inteins, Valuable Genetic Elements in Molecular Biology and Biotechnology.” *Applied Microbiology and Biotechnology*. https://doi.org/10.1007/S00253-010-2628-X.

Emsley, P., B. Lohkamp, W. G. Scott, and K. Cowtan. 2010. “Features and Development of Coot.” *Acta Crystallographica Section D: Biological Crystallography*. https://doi.org/10.1107/S0907444910007493.

Essmann, Ulrich, Lalith Perera, Max L. Berkowitz, Tom Darden, Hsing Lee, and Lee G. Pedersen. 1995. “A Smooth Particle Mesh Ewald Method.” *The Journal of Chemical Physics*. https://doi.org/10.1063/1.470117.

Frame, Margaret C., Hitesh Patel, Bryan Serrels, Daniel Lietha, and Michael J. Eck. 2010. “The FERM Domain: Organizing the Structure and Function of FAK.” *Nature Reviews Molecular Cell Biology*. https://doi.org/10.1038/nrm2996.

Gogarten, J. Peter, Alireza G. Senejani, Olga Zhaxybayeva, Lorraine Olendzenski, and Elena Hilario. 2002. “Inteins: Structure, Function, and Evolution.” *Annual Review of Microbiology*. https://doi.org/10.1146/annurev.micro.56.012302.160741.

Han, Jung Hoon, Sarah Batey, Adrian A. Nickson, Sarah A. Teichmann, and Jane Clarke. 2007. “The Folding and Evolution of Multidomain Proteins.” *Nature Reviews Molecular Cell Biology*. https://doi.org/10.1038/nrm2144.

Harrison, Joseph S, Tim M Jacobs, Kevin Houlihan, Koenraad Van Doorslaer, and Brian Kuhlman. 2016. “UbSRD: The Ubiquitin Structural Relational Database.” *Journal of Molecular Biology*. https://doi.org/10.1016/j.jmb.2015.09.011.

Hicke, Linda, Heidi L. Schubert, and Christopher P. Hill. 2005. “Ubiquitin-Binding Domains.” *Nature Reviews Molecular Cell Biology*. https://doi.org/10.1038/nrm1701.

Holm, L., and P. Rosenstrom. 2010. “Dali Server: Conservation Mapping in 3D.” *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkq366.

Humphrey, W, A Dalke, and K Schulten. 1996. “VMD: Visual Molecular Dynamics.” *Journal of Molecular Graphics*. http://www.ncbi.nlm.nih.gov/pubmed/8744570.

Hurley, James H, Sangho Lee, and Gali Prag. 2006. “Ubiquitin-Binding Domains.” *The Biochemical Journal*. https://doi.org/10.1042/BJ20061138.

Jorgensen, William L., Jayaraman Chandrasekhar, Jeffry D. Madura, Roger W. Impey, and Michael L. Klein. 1983. “Comparison of Simple Potential Functions for Simulating Liquid Water.” *The Journal of Chemical Physics*. https://doi.org/10.1063/1.445869.

Kabsch, Wolfgang, Brünger A. T., Diederichs K., Karplus P. A., Diederichs K., McSweeney S., Ravelli R. B. G., et al. 2010. “*XDS*.” *Acta Crystallographica Section D Biological Crystallography*. https://doi.org/10.1107/S0907444909047337.

Kamitani, Tetsu, Katsumi Kito, Hung P. Nguyen, and Edward T.H. Yeh. 1997. “Characterization of NEDD8, a Developmentally down-Regulated Ubiquitin- like Protein.” *Journal of Biological Chemistry*.

https://doi.org/10.1074/jbc.272.45.28557.

Katoh, K., and D. M. Standley. 2013. “MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability.” *Molecular Biology and Evolution*. https://doi.org/10.1093/molbev/mst010.

Khan, Amir R., and Micael N. G. James. 2008. “Molecular Mechanisms for the Conversion of Zymogens to Active Proteolytic Enzymes.” *Protein Science*. https://doi.org/10.1002/pro.5560070401.

Kniss, Andreas, Denise Schuetz, Sina Kazemi, Lukas Pluska, Philipp E. Spindler, Vladimir V. Rogov, Koraljka Husnjak, et al. 2018. “Chain Assembly and Disassembly Processes Differently Affect the Conformational Space of Ubiquitin Chains.” *Structure*. https://doi.org/10.1016/j.str.2017.12.011.

Komander, David, and Michael Rape. 2012. “The Ubiquitin Code.” *Annual Review of Biochemistry*. https://doi.org/10.1146/annurev-biochem-060310-170328.

Lew, Belinda M., Kenneth V. Mills, and Henry Paulus. 1998. “Protein Splicing in Vitro with a Semisynthetic Two-Component Minimal Intein.” *Journal of Biological Chemistry*. https://doi.org/10.1074/jbc.273.26.15887.

Lipkowitz, Stanley, and Allan M. Weissman. 2011. “RINGs of Good and Evil: RING Finger Ubiquitin Ligases at the Crossroads of Tumour Suppression and Oncogenesis.” *Nature Reviews Cancer*. https://doi.org/10.1038/nrc3120.

Liu, Dongsheng, and David Cowburn. 2017. “Segmental Isotopic Labeling of Proteins for NMR Study Using Intein Technology.” In

Methods in Molecular Biology. https://doi.org/10.1007/978-1-4939-6451-2_9.

Mills, Kenneth V., and Henry Paulus. 2001. "Reversible Inhibition of Protein Splicing by Zinc Ion." *Journal of Biological Chemistry*. https://doi.org/10.1074/jbc.M011149200.

Muona, M, A S Aranko, V Raulinaitis, and H Iwai. 2010. "Segmental Isotopic Labeling of Multi-Domain and Fusion Proteins by Protein Trans-Splicing in Vivo and in Vitro." *Nat Protoc*. https://doi.org/10.1038/nprot.2009.240.

Murshudov, Garib N., Pavol Skubák, Andrey A. Lebedev, Navraj S. Pannu, Roberto A. Steiner, Robert A. Nicholls, Martyn D. Winn, Fei Long, and Alexei A. Vagin. 2011. "REFMAC5 for the Refinement of Macromolecular Crystal Structures." *Acta Crystallographica Section D: Biological Crystallography*. https://doi.org/10.1107/S0907444911001314.

Nakayama, Keiichi I., and Keiko Nakayama. 2006. "Ubiquitin Ligases: Cell-Cycle Control and Cancer." *Nature Reviews Cancer*. https://doi.org/10.1038/nrc1881.

Nichols, Nicole Magnasco, Jack S. Benner, Deana D. Martin, and Thomas C. Evans. 2003. "Zinc Ion Effects on Individual *Ssp* DnaE Intein Splicing Steps: Regulating Pathway Progression." *Biochemistry*. https://doi.org/10.1021/bi020679e.

O'Sullivan, Orla, Karsten Suhre, Chantal Abergel, Desmond G Higgins, and Cédric Notredame. 2004. "3DCoffee: Combining Protein Sequences and Structures within Multiple Sequence Alignments." *Journal of Molecular Biology*.

https://doi.org/10.1016/j.jmb.2004.04.058.

Pavankumar, Theetha. 2018. “Inteins: Localized Distribution, Gene Regulation, and Protein Engineering for Biological Applications.” *Microorganisms*. https://doi.org/10.3390/microorganisms6010019.

Perler, Francine B. 1998. “Protein Splicing of Inteins and Hedgehog Autoproteolysis: Structure, Function, and Evolution.” *Cell*. https://doi.org/10.1016/S0092-8674(00)80892-2.

Pettersen, Eric F., Thomas D. Goddard, Conrad C. Huang, Gregory S. Couch, Daniel M. Greenblatt, Elaine C. Meng, and Thomas E. Ferrin. 2004. “UCSF Chimera?A Visualization System for Exploratory Research and Analysis.” *Journal of Computational Chemistry*. https://doi.org/10.1002/jcc.20084.

Pietrokovski, S. 2001. “Intein Spread and Extinction in Evolution.” *Trends in Genetics*. https://doi.org/10.1016/s0168-9525(01)02365-4.

Reily, Colin, Tyler J. Stewart, Matthew B. Renfrow, and Jan Novak. 2019. “Glycosylation in Health and Disease.” *Nature Reviews Nephrology*. https://doi.org/10.1038/s41581-019-0129-4.

Roy, Ambrish, Alper Kucukural, and Yang Zhang. 2010. “I-TASSER: A Unified Platform for Automated Protein Structure and Function Prediction.” *Nature Protocols*. https://doi.org/10.1038/nprot.2010.5.

Saha, Suvrajit, Anupama Ambika Anilkumar, and Satyajit Mayor. 2016. “GPI-Anchored Protein Organization and Dynamics at the Cell Surface.” *Journal of Lipid Research*. https://doi.org/10.1194/jlr.R062885.

Sasaki, Atsuo T., Arkaitz Carracedo, Jason W. Locasale, Dimitrios Anastasiou, Koh Takeuchi, Emily Rose Kahoud, Sasson Haviv, John

M. Asara, Pier Paolo Pandolfi, and Lewis C. Cantley. 2011. "Ubiquitination of K-Ras Enhances Activation and Facilitates Binding to Select Downstream Effectors." *Science Signaling*. https://doi.org/10.1126/scisignal.2001518.

Sauvé, Véronique, Asparouh Lilov, Marjan Seirafi, Marta Vranas, Shafqat Rasool, Guennadi Kozlov, Tara Sprules, Jimin Wang, Jean-François Trempe, and Kalle Gehring. 2015. "A Ubl/Ubiquitin Switch in the Activation of Parkin." *The EMBO Journal*. https://doi.org/10.15252/embj.201592237.

Shah, Neel H, and Tom W Muir. 2014. "Inteins: Nature's Gift to Protein Chemists." *Chemical Science*. https://doi.org/10.1039/C3SC52951G.

Shao, Yang, and Stephen B.H. Kent. 1997. "Protein Splicing: Occurrence, Mechanisms and Related Phenomena." *Chemistry and Biology*. https://doi.org/10.1016/S1074-5521(97)90287-8.

Soucy, Shannon M., Matthew S. Fullmer, R. Thane Papke, and Johann Peter Gogarten. 2014. "Inteins as Indicators of Gene Flow in the Halobacteria." *Frontiers in Microbiology*. https://doi.org/10.3389/fmicb.2014.00299.

Swatek, Kirby N., and David Komander. 2016. "Ubiquitin Modifications." *Cell Research*. https://doi.org/10.1038/cr.2016.39.

Taherbhoy, Asad M., Brenda A. Schulman, and Stephen E. Kaiser. 2012. "Ubiquitin-like Modifiers." *Essays In Biochemistry*. http://essays.biochemistry.org/content/52/51.long.

Terwilliger, Thomas C., Paul D. Adams, Randy J. Read, Airlie J. McCoy, Nigel W. Moriarty, Ralf W. Grosse-Kunstleve, Pavel V. Afonine, Peter H. Zwart, and Li Wei Hung. 2009. "Decision-Making in

Structure Solution Using Bayesian Estimates of Map Quality: The PHENIX AutoSol Wizard." *Acta Crystallographica Section D: Biological Crystallography*. https://doi.org/10.1107/S0907444909012098.

Terwilliger, Thomas C., Ralf W. Grosse-Kunstleve, Pavel V. Afonine, Nigel W. Moriarty, Peter H. Zwart, Li Wei Hung, Randy J. Read, and Paul D. Adams. 2007. "Iterative Model Building, Structure Refinement and Density Modification with the PHENIX AutoBuild Wizard." In *Acta Crystallographica Section D: Biological Crystallography*. https://doi.org/10.1107/S090744490705024X.

Topilina, Natalya I., Olga Novikova, Matthew Stanger, Nilesh K. Banavali, and Marlene Belfort. 2015a. "Post-Translational Environmental Switch of RadA Activity by Extein-Intein Interactions in Protein Splicing." *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkv612.

Vagin, Alexei, and Alexei Teplyakov. 2010. "Molecular Replacement with *MOLREP*." *Acta Crystallographica Section D Biological Crystallography*. https://doi.org/10.1107/S0907444909042589.

Vanommeslaeghe, K, E Hatcher, C Acharya, S Kundu, S Zhong, J Shim, E Darian, et al. 2010. "CHARMM General Force Field: A Force Field for Drug-like Molecules Compatible with the CHARMM All-Atom Additive Biological Force Fields." *Journal of Computational Chemistry*. https://doi.org/10.1002/jcc.21367.

Verlet, Loup. 1967. "Computer 'Experiments' on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules." *Physical Review*. https://doi.org/10.1103/PhysRev.159.98.

Wood, David W., Wei Wu, Georges Belfort, Victoria Derbyshire, and Marlene Belfort. 1999. “A Genetic System Yields Self-Cleaving Inteins for Bioseparations.” *Nature Biotechnology*. https://doi.org/10.1038/12879.

Zuin, Alice, Marta Isasa, and Bernat Crosas. 2014. “Ubiquitin Signaling: Extreme Conservation as a Source of Diversity.” *Cells*. https://doi.org/10.3390/cells3030690.